DOCUMENT RESUME

ED 318 074 CS 507 148

AUTHOR	Pisoni, David B.
TITLE	Research on Speech Perception. Progress Report No. 15.
INSTITUTION	Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY	Air Force Armstrong Aerospace Medical Research Lab, Wright-Patterson AFB, OH.; National Institutes of Health (DHHS), Bethesda, Md.; National Science Foundation, Washington, D.C.
PUB DATE	89
CONTRACT	AF-F-33615-86-C-0549
GRANT	DC-00012-11; DC-00111-13; IRI-86-17847
NOTE	519p.; For other reports in this series, see CS 507 123-129.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF02/PC21 Plus Postage.
DESCRIPTORS	Acoustic Phonetics; Auditory Discrimination; *Auditory Perception; Communication Research; Infants; Language Processing; Language Research; Linguistics; *Speech; Speech Synthesizers
IDENTIFIERS	Indiana University Bloomington; *Speech Perception; Speech Research; Theory Development

ABSTRACT

Summarizing research activities in 1989, this is the fifteenth annual report of research on speech perception, analysis, synthesis, and recognition conducted in the Speech Research Laboratory of the Department of Psychology at Indiana University. The report contains the following 21 articles: "Perceptual Learning of Nonnative Speech Contrasts: Implications for Theories of Speech Perception" (D. B. Pisoni and others); "Modes of Processing Speech and Nonspeech Signals" (D. B. Pisoni); "Comprehension of Synthetic Speech Produced by Rule" (J. V. Ralston and others); "Effects of Talker Variability on Speech Perception by 2-Month-Old Infants" (P. W. Jusczyk and others); "Neighborhood Density Effects for High Frequency Words: Evidence for Activation-Based Models of Word Recognition" (S. D. Goldinger); "Contrast and Normalization in Vowel Perception" (K. Johnson); "Coronals and the Phonotactics of Nonadjacent Consonants in English" (S. A. Davis); "Position of the Maximum Amplitude as a Perceptual Stress Cue in English: Work in Progress" (D. M. Behne); "A Comparison of the First and Second Formants of Vowels Common to English and French" (D. M. Behne); "Age Differences in Spoken Word Identification: Effects of Lexical Density and Semantic Context" (T. S. Bell); "Movement Dynamics and the Nature of Errors in Tongue Twisters: An Observation and Research Proposal" (S. D. Goldinger); "Lexical Neighborhoods in Speech Production: A First Report" (S. D. Goldinger and W. V. Summers); "On the Perceptual Representation of Vowel Categories" (K. Johnson); "Glottal Effects of LPC Estimation of F1" (K. Johnson); "Stress-Class in Isolated Phrases and Sentence Contexts" (K. Johnson and M. S. Cluff); "Final Report to the NTSB on the Speech Produced by the Captain of the Exxon Valdez" (K. Johnson and others); "Effects of Talker Familiarity on Serial Recall of Spoken Word Lists" (N. L. Lightfoot); "Inhibition or Facilitation? An Investigation of Form-Based Priming and Response Bias in Spoken Word Recognition" (J. K. Marcario and S. D. Goldinger); "Talker Variability and Spoken Word Recognition: A Developmental Study" (B. R. Oliver); "Effects of Cognitive Workload on Speech Production: Acoustic Analyses" (W. V. Summers and others); and "Current Computer Facilities in the Speech Research Laboratory" (R. H. Bernacki and others). Lists of publications and of laboratory staff and personnel conclude the report. (SR)

ED318074

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 15
(1989)

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. DC-00111-13

National Institutes of Health
Training Grant No. DC-00012-11

National Science Foundation
Research Grant No. IRI-86-17847

and

U.S. Air Force
Armstrong Aerospace Medical Research Laboratory
Contract No. AF-F-33615-86-C-0549

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
DAVID B. PISONI, PhD
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS507148

BEST COPY AVAILABLE

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 15
(1989)

David B. Pisoni, Ph.D.
Principal Investigator

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

Research Supported by:

Department of Health and Human Services
U. S. Public Health Service

National Institutes of Health
Research Grant No. DC-00111-13

National Institutes of Health
Training Grant No. DC-00012-11

National Science Foundation
Research Grant No. IRI-86-17847

and

U. S. Air Force
Armstrong Aerospace Medical Research Laboratory
Contract No. AF-F-33615-86-C-0549

RESEARCH ON SPEECH PERCEPTION Progress Report No. 15 (1989)

Table of Contents

Introduction iv

I. Extended Manuscripts 1

Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception;
David B. Pisoni, John S. Logan and Scott E. Lively 3

Modes of processing speech and nonspeech signals;
David B. Pisoni 59

Comprehension of synthetic speech produced by rule;
James V. Ralston, David B. Pisoni and John W. Mullennix 77

Effects of talker variability on speech perception by 2-month-old infants;
Peter W. Jusczyk, David B. Pisoni and John W. Mullennix 133

Neighborhood density effects for high frequency words: Evidence for activation-based models of word recognition;
Stephen D. Goldinger 163

Contrast and normalization in vowel perception;
Keith Johnson 187

Coronals and the phonotactics of nonadjacent consonants in English;
Stuart A. Davis 227

II. Short Reports and Work-in-Progress 241

Position of the maximum amplitude as a perceptual stress cue in English: Work in progress;
Dawn M. Behne 243

A comparison of the first and second formants of vowels common to English and French;
Dawn M. Behne 269

Age differences in spoken word identification: Effects of lexical density and semantic context;
Theodore S. Bell 283

Movement dynamics and the nature of errors in tongue twisters: An observation and research proposal;
Stephen D. Goldinger 303

Lexical neighborhoods in speech production: A first report;
Stephen D. Goldinger and W. Van Summers 331

On the perceptual representation of vowel categories;
Keith Johnson 343

Glottal effects on LPC estimation of F1;
Keith Johnson 359

Stress-clash in isolated phrases and sentence contexts;
Keith Johnson and Michael S. Cluff 373

Final report to the NTSB on the speech produced by the Captain of the Exxon Valdez;
Keith Johnson, David B. Pisoni and Robert H. Bernacki 389

Effects of talker familiarity on serial recall of spoken word lists;
Nancy L. Lightfoot 419

Inhibition or facilitation? An investigation of form-based priming and response bias in spoken word recognition;
Joanne K. Marcario and Stephen D. Goldinger 445

Talker variability and spoken word recognition: A developmental study;
Brigette R. Oliver 471

Effects of cognitive workload on speech production: Acoustic analyses;
W. Van Summers, David B. Pisoni and Robert H. Bernacki 485

III. Instrumentation and Software Development 503

Current computer facilities in the speech research laboratory;
Robert H. Bernacki, Dennis M. Feaster, Luis R. Hernandez and
Jerry C. Forshee 505

IV. Publications 513

V. SRL Laboratory Staff and Personnel 517

6

This is the fifteenth annual report summarizing the research activities on speech perception, analysis, synthesis, and recognition carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various research activities over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software support when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating our own research.

We are distributing reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of speech processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis, synthesis, and recognition and, therefore, we would be grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
USA
(812) 855-1155, 855-1768
E-mail (BITNET) "PISONI@IUBACS"
E-mail (INTERNET) "PISONI@UCS.INDIANA.EDU"

Copies of this report are being sent primarily to libraries and specific research institutions rather than individual scientists. Because of the rising costs of publication and printing, it is not possible to provide multiple copies of this report to people at the same institution or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address.

The information contained in the report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.

I. EXTENDED MANUSCRIPTS

Perceptual Learning of Nonnative Speech Contrasts: Implications for Theories of Speech Perception

David B. Pisoni, John S. Logan and Scott E. Lively[1]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] Chapter to appear in H.C. Nusbaum and J. Goodman (Eds.), *Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words*. Cambridge, MIT Press, 1990 (In Press). Preparation of this chapter was supported by NIDCD Research Grant R01 DC00111-13 to Indiana University in Bloomington. We thank Daniel Dinnsen, Judith Gierut and Robert Nosofsky for suggestions and advice at various stages of this work.

Perceptual Learning of Nonnative Speech Contrasts: Implications for Theories of Speech Perception

For many years, there has been a consensus among investigators working in the field of speech perception that the linguistic environment exerts a very profound and often quite permanent effect on an individual's ability to identify and discriminate speech sounds. The first report of categorical perception by Liberman and his colleagues at Haskins Laboratories (Liberman, Harris, Hoffman & Griffith, 1957) and the subsequent cross-language studies of voicing by Lisker and Abramson (1964, 1967) provided very convincing evidence for the important role of perceptual learning in speech perception. These initial studies, and many others since then, have demonstrated that the effects of perceptual learning are long-lasting and often produce permanent and seemingly non-reversible changes in the speech perception abilities of adults. Indeed, most attempts to selectively modify speech perception abilities using short-term laboratory training techniques have been generally unsuccessful (Strange & Jenkins, 1978; Strange & Dittman, 1984). The failure of these earlier training studies to produce robust changes in speech perception has been interpreted by some researchers as strong support for the proposal that during development the underlying neural mechanisms used in speech perception become very finely tuned to only the distinctive sound contrasts used in the linguistic environment, and they cannot be selectively modified or "retuned" very easily in mature adults in a short period of time (Strange & Jenkins, 1978; Eimas, 1975, 1978).

The present chapter is concerned with several general issues surrounding perceptual learning in speech perception. While our major interest will be focused primarily on learning of nonnative speech contrasts in mature adults, much of what we have to say here will also be relevant to several other issues dealing with current theoretical accounts of speech perception, perceptual development and units of perceptual analysis. Central to our discussion is a concern for the nature of the changes that take place when the sound system of a language is acquired in development. In particular, we are interested in what happens to a listener's perceptual abilities when he/she acquires a native language. What happens to a listener's ability to identify and discriminate speech contrasts that are not present in the language-learning environment? Are the listener's perceptual abilities permanently "lost" because the neural mechanisms have atrophied due to lack of stimulation during development, or are they simply realigned and only temporarily modified due to changes in selective attention? Despite the existence of several recent studies in the published literature demonstrating that under certain experimental conditions listeners can be trained to perceive and discriminate very fine phonetic details, many researchers continue to maintain and proliferate the view that the effects of linguistic experience on speech perception are difficult, if not impossible, to overcome and modify in a short period of time. The quotations below should give the reader sufficient evidence of the pervasiveness of these views in the literature:

Thus, for adults learning a foreign language, modification of phonetic perception appears to be slow and effortful, and is characterized by considerable variability among individuals. (Strange & Dittman, 1984, p. 132)

These difficulties with non-native speech contrasts may indicate that certain distinctions are extremely difficult for adults to learn, or even that adults cannot learn to make certain distinctions in a linguistically meaningful manner. (Jamieson & Morosan, 1986, p. 206)

An English-speaking adult, for example, has difficulty perceiving the difference between the two /p/ phones that are used in Thai (Lisker & Abramson, 1970). So too, a Japanese-speaking adult initially cannot distinguish between the English /ra/ and /la/, because Japanese uses a single phoneme intermediate between the two English phonemes. (Werker, 1989, pp. 54-55)

The language environment modifies the speech perception abilities found in early development. In particular, adults have difficulty perceiving many phonetic contrasts that young infants discriminate. (Best, McRoberts, & Sithole, 1988, p.345)

The extreme difficulty that Japanese adults demonstrate in learning to differentiate these phonemes illustrates the profound effect that first language learning has in modifying what may be innate discriminative processes. (Sheldon & Strange, 1982, p. 254)

As noted previously, in the absence of early experience with a language in which /l/ and /r/ are contrastive, many native speakers of Japanese are unable to distinguish utterances which contain English /l/ and /r/ in either labeling or discrimination tasks which focus on the /l/-/r/ distinction. (Mann, 1986, p. 174)

Everyone knows, of course, that native speakers of Japanese have trouble pronouncing /r/ and /l/. What is not so well known is that native speakers of Japanese do not *hear* the difference between /r/ and /l/ either. That is, when a speaker of American English says *rock* or *lock*, the Japanese cannot tell you which one was said; in fact, they may not be able to tell the syllables were different from each other. (Jenkins, 1989, p. 481)

Role of Early Experience in Perceptual Development

Most current theories of speech perception are vague and it has often been difficult to generate specific, testable experimental hypotheses (see Pisoni, 1978; Pisoni & Luce, 1986). A detailed examination of these theories will reveal that none of them currently incorporate mechanisms or procedures to deal with developmental change or the effects of the linguistic environment on speech perception. All contemporary theories of speech perception are concerned with the mature adult listener who is presumably in the end-state of development. We believe this is an unfortunate state of affairs because theories of speech perception should not only characterize the perceptual abilities of the mature listener but also should provide some principled account of how these abilities developed over time and how they come to be modified selectively by the linguistic environment. Some initial efforts have already been made by Jusczyk (1985; 1986) and Studdert-Kennedy (1986; 1987) to deal with problems of development in speech perception in young infants, but many researchers continue to focus their attention and efforts exclusively on the mature adult listener with little, if any, concern for how these abilities developed. The results we describe below make it very clear that current theories of speech perception will have to be modified in several ways to incorporate principles of developmental change to account for how mature adults can acquire new linguistic contrasts that are not present in their native language.

To place our work in a developmental framework, we first consider some possible interactions between genetic and experiential factors in perceptual development. These ideas were initially formulated by Aslin and Pisoni (1980) in an attempt to deal with the ontogeny of infant speech perception. An examination of the literature on infant speech perception revealed a complex set of interactions among genetic and experiential factors in development. Following observations of researchers working in visual system development and suggestions made by Gottlieb (1981), it became clear to us that a simple dichotomy between nativist and empiricist views of development was inadequate to account for the interactions that underlie normal perceptual development. To deal with these interactions, Aslin and Pisoni (1980) proposed an account of the possible roles that early experience can play in the development of speech perception. These alternatives are shown in Figure 1.

Insert Figure 1 about here

First, a perceptual ability may be present at birth but require certain specific types of early experience to *maintain* the integrity of that ability. The absence or degradation of the prerequisite early experience can result in either a partial or complete loss of the perceptual ability, a loss that may be irreversible despite subsequent experience at a later point in development.

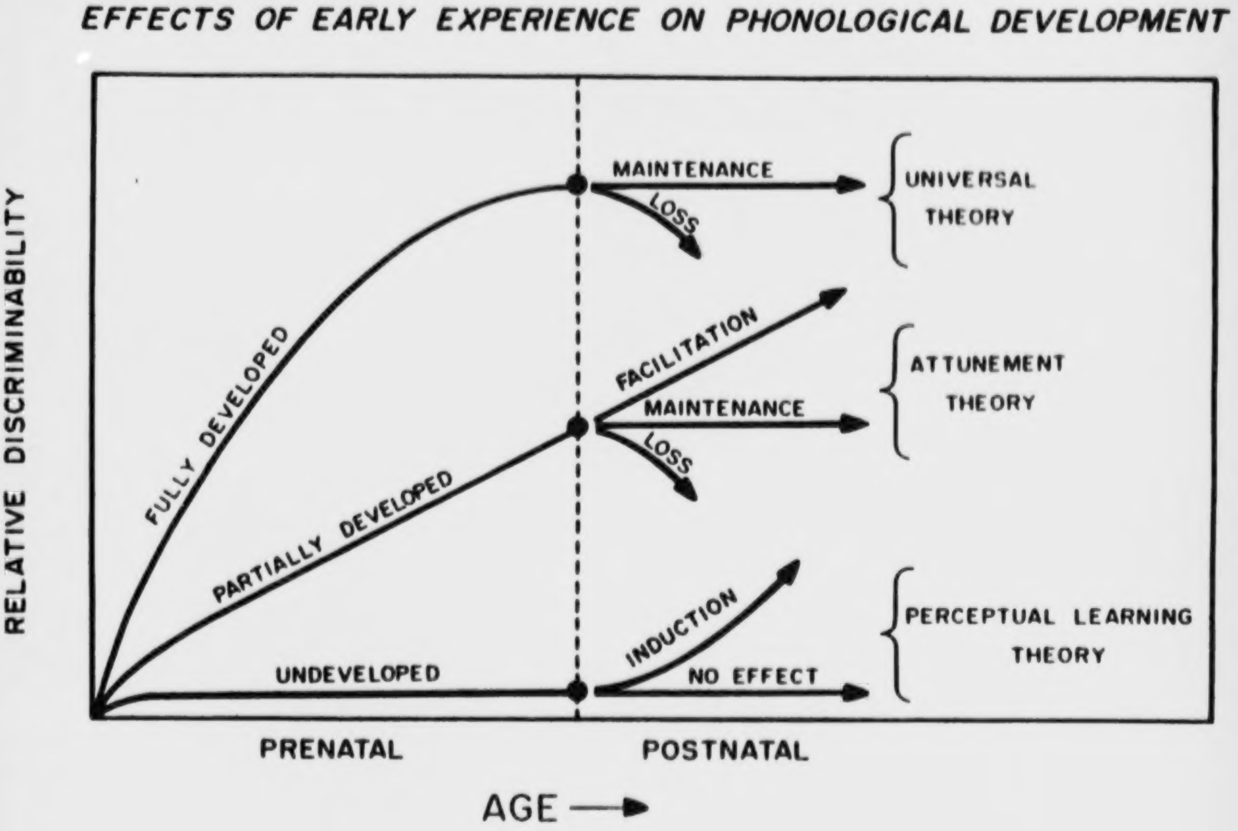

EFFECTS OF EARLY EXPERIENCE ON PHONOLOGICAL DEVELOPMENT
RELATIVE DISCRIMINABILITY
FULLY DEVELOPED
PARTIALLY DEVELOPED
UNDEVELOPED
MAINTENANCE
LOSS
UNIVERSAL THEORY
FACILITATION
MAINTENANCE
LOSS
ATTUNEMENT THEORY
INDUCTION
NO EFFECT
PERCEPTUAL LEARNING THEORY
PRENATAL
POSTNATAL
AGE

Figure 1. Illustration of the major roles that early experience can play in modifying the perception of speech sound contrasts. Three general classes of perceptual theories are shown here: Universal Theory, Attunement Theory and Perceptual Learning Theory (from Aslin & Pisoni, 1980).

Second, a perceptual ability may be only partially developed at birth and require specific types of early experience to *facilitate* or attune the further development of that ability. The absence of early experience with these critical stimuli, which may serve a facilitating function during "sensitive periods" in development, could result either in the absence of any further development or a loss of that ability when compared to its level at birth.

Third, a perceptual ability may be absent at birth, and its development may depend on a process of induction based on specific early experiences of the organism in the environment. The presence of a particular ability, then, would depend to a large extent on the presence of a particular type of early experience. Thus, specific kinds of early experience are necessary for the subsequent development and maintenance of a particular preference or tendency.

Finally, early experience may, of course, exert no role at all in the development of a particular perceptual ability. That is, the ability may be either present or absent at birth and it may remain, decline or improve in the absence of any specific type of early experience. Absence of experiential effects is difficult to identify and often leads to unwarranted conclusions, especially those that assume that an induction process might be operative. For example, it has been common for researchers to argue that if an ability is absent at birth, the ability must have been learned (see Eilers et al., 1979). In terms of the conceptual framework outlined here, this could be an instance of induction. However, it is quite possible that the ability simply unfolded developmentally according to a genetically specified maturational schedule – a schedule that required no particular type of early experience in the environment. This unfolding of an ability may be thought of as adhering to the general class of maturational theories of development.

The complexity of these numerous options – maintenance, facilitation, induction, and maturation – and their possible interactions suggest that researchers should be cautious about drawing any strong conclusions about the developmental course of specific perceptual abilities. In the context of speech perception, we believe that several conclusions that are now quite prominent in the literature about the role of early experience may have been premature and possibly even unjustified given the findings described below (see Pisoni et al., 1982). In order to clarify the roles of early experience in the development of speech perception and to put these ideas into a somewhat broader theoretical context, we briefly consider four general classes of theories of perceptual development. These theories are: Universal Theory, Attunement Theory, Perceptual Learning Theory and Maturational Theory.

Universal Theory assumes that at birth infants are capable of discriminating all the possible phonetic contrasts that may be used distinctively in any natural language. According to this view, early experience functions to maintain the ability to discriminate phonetically relevant distinctions – those distinctions actually used in the language-learning environment of the infant. However, the absence of exposure to phonetically-irrelevant contrasts may result in a selective loss of the abilities to discriminate those specific contrasts. The perceptual mechanisms responsible for this loss of sensitivity may be neural, attentional or both.

Universal Theory makes several specific predictions concerning the possible reacquisition of the lost discriminative abilities in mature adults, a topic that is relevant to the goals of the present chapter. For example, if a child is exposed to some phonetic contrasts that are not phonologically distinctive in the language because of allophonic variation, it may be possible to discriminate these quite easily because the underlying perceptual mechanisms are still operative. In contrast, lack of exposure during development may produce a permanent loss or attenuation in the ability that cannot be overcome by training or exposure. In either case, it becomes important to determine if the loss has a sensory-perceptual basis or if it is due primarily to changes in selective attention.

Attunement Theory assumes that at birth all infants are capable of discriminating at least some of the possible phonetic contrasts present in the world's languages, but that the infant's discriminative capacities are incompletely developed, quite broadly tuned or both. According to this view, early experience functions to "align" and/or sharpen these partially developed discriminative abilities. Phonetically-relevant contrasts in the language-learning environment become more finely tuned with experience and phonetically-irrelevant contrasts either remain broadly tuned or become attenuated in the absence of specific environmental stimulation.

In contrast with the other two views, *Perceptual Learning Theory* assumes that the ability to discriminate any particular phonetic contrast is highly dependent on specific early experience with that sound contrast in the language-learning environment. The rate of development could be very fast or very slow depending on the relative importance of the phonetic contrasts during early life, the relative psychophysical discriminability of the acoustic attributes compared with other phonetic contrasts, and the attentional state of the infant. According to this view, however, phonetically-irrelevant contrasts would initially never be discriminated better than the phonetically-relevant ones that are present in the language-learning environment.

Finally, *Maturational Theory* assumes that the ability to discriminate a particular phonetic contrast is independent of any specific early experience and simply "unfolds" according to a predetermined developmental schedule. According to this view, all possible phonetic contrasts would be discriminated equally well irrespective of the language-learning environment, although the age at which specific phonetic contrasts could be discriminated would be dependent on the developmental level of the underlying sensory mechanisms. For example, if young infants did not show sensitivity to high frequencies until later in development, one would not expect them to discriminate phonetic contrasts that were differentiated on the basis of high frequency information at birth.

These four general classes of theories make specific predictions about the developmental course of speech perception and the underlying perceptual abilities of mature listeners. It is important to point out here that probably no single class of theories will uniquely account for the development of all speech contrasts. Rather, it may be the case that some hybrid parallel version of the theories provides the best overall description of the perception of specific classes of speech sounds. In fact, this view of parallel developmental processes appears to be well supported by current research findings. Examples of Attunement Theory have been found by Werker (1989) who has reported a series of studies showing evidence for excellent phonetic sensitivity in young infants followed by a developmental decline of these perceptual abilities in adulthood. For the phonetic contrasts she studied, the decline in phonetic sensitivity and perceptual reorganization occurred between six and twelve months of age and appeared to be a function of specific language experience. An example of Universal Theory was reported by Best, McRoberts, and Sithole (1988) who studied the perception of Zulu clicks by English adults and infants. In contrast to Werker's findings, they found that both infants *and* adults were able to discriminate these sounds despite the lack of experience hearing clicks spoken in their language learning environment. Considering the potential complex interactions between genetic and experiential factors in perceptual development described above, an important long-term goal of research in speech perception becomes the investigation of the development of as many phonetic contrasts in language as possible in order to understand the underlying perceptual mechanisms and the way they become selectively modified by early experience (Aslin, 1981; Aslin, 1985; Werker, 1989; Best et al., 1988).

In the sections below, we consider two phonetic contrasts that have occupied the attention of speech researchers over the last few years. The first contrast is the voicing distinction in initial stop consonants. The second contrast is the distinction between /r/ and /l/. Both phonetic contrasts have played an important role in recent theorizing about the effects of early experience on speech perception and both contrasts have been used in studies that were designed to selectively modify the perceptual analysis of these sounds in mature adult listeners. Because these two contrasts have quite different acoustic correlates and phonological properties in different languages, they are ideal candidates to consider in studies of perceptual learning.

Perception of Voicing Contrasts in Stop Consonants

Over the last twenty years numerous studies employing synthetically produced speech stimuli have investigated the perception of *voice-onset-time* (VOT) in human adults, human infants, chinchillas and monkeys. These developmental and cross-species comparisons have been undertaken to study the potential interactions between genetic predispositions and experiential factors in speech perception. The results of these diverse studies have shown the combined influence of both factors. First, linguistic experience has been shown to have a substantial effect on speech perception, particularly in human adults exposed to different language-learning environments (Lisker & Abramson, 1964). Subjects identify and discriminate speech sounds with reference to the linguistic categories of their language. Second, basic sensory and psychophysical constraints on auditory system function seem to affect perception of both speech and nonspeech control signals in similar ways. For example, the perception of voicing in stop consonants apparently requires the analysis of a temporal relation between laryngeal and supralaryngeal events (Pisoni, 1977). Basic constraints on auditory perception appear to play an important role in defining the inventory of acoustic correlates for distinctive features used in speech (Stevens, 1972; 1980). This inventory is then modified and reorganized selectively by the speakers and hearers in a language-learning environment.

The results of the earliest cross-language experiments on the perception and production of VOT by Lisker and Abramson (1964, 1967) confirmed that the linguistic environment exerts a profound influence on the ability to produce and perceive voicing differences in initial stop consonants. They examined the voicing and aspiration differences among stops produced by native speakers from eleven diverse languages and were able to identify three primary modes of voicing: (1) a lead mode in which voicing onset precedes the release from stop closure, (2) a short-lag mode in which voicing onset is roughly simultaneous with release from stop closure, and (3) a long-lag mode in which voicing onset occurs substantially after the release. In addition to measurements of VOT in the production of stop contrasts, Lisker and Abramson (1967) and Abramson and Lisker (1970) also carried out several perceptual experiments using synthetically produced speech stimuli that differed in VOT. The results of these perceptual experiments demonstrated that subjects from different linguistic backgrounds identify and discriminate speech stimuli in terms of the distinctive phonological categories used in their language.

Insert Figure 2 about here

Figure 2 shows the identification functions for native speakers of English, Thai and Spanish for three series of synthetic stimuli differing in VOT. The functions display perceptual boundaries at either one or two locations along the VOT continuum, corresponding to the presence of two or three voicing categories. The discrimination functions which are not shown here revealed discontinuities along the stimulus continuum with peaks located at the cross-over points separating perceptual categories in identification. The correspondence of heightened discrimination at the category boundaries combined with relatively poor discrimination within perceptual categories demonstrated that subjects could discriminate between stimuli only as well as they could identify them as different on an absolute basis and suggested that the perceptual categories are determined, in large part, by the linguistic experience of the listener.

The subjects in these early perceptual experiments, as well as those used in more recent studies, appeared to have a great deal of difficulty in identifying and subsequently discriminating between stimuli that were *not* distinctive in their native language. The failure of adults to perceive non-native distinctions in voicing has been interpreted by a number of investigators as support for the view that linguistic experience exerts a profound and lasting

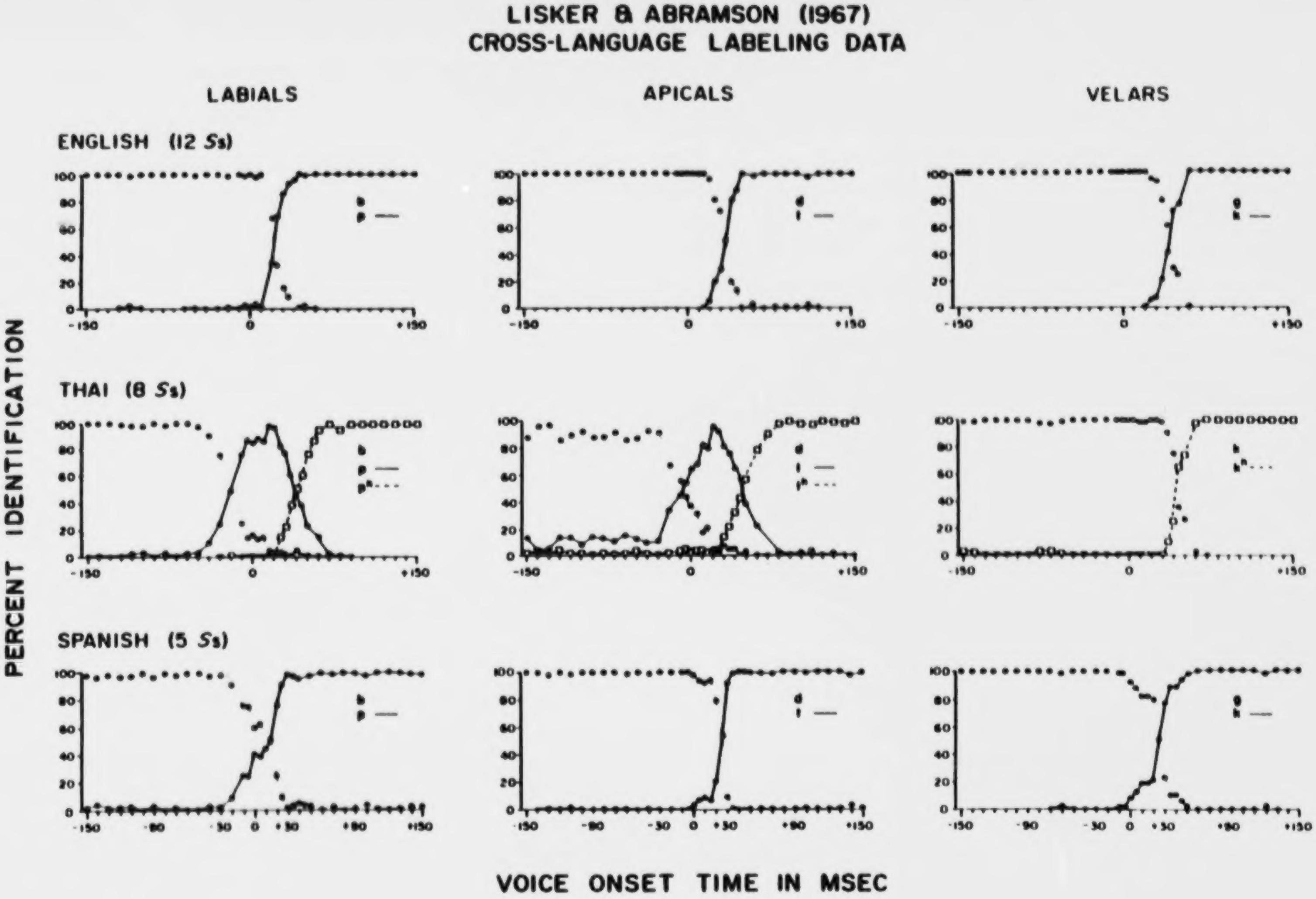


Figure 2. Adult identification functions for synthetic labial, apical and velar stop consonants differing in voice onset time (VOT). The data were obtained from native speakers of English (N = 12), Thai (N = 8) and Spanish (N = 5). (Adapted from Lisker and Abramson, 1967).

effect on an individual's ability to discriminate speech stimuli. Indeed, based on his work with young infants, Eimas (1978) has even suggested that the neural mechanisms mediating VOT perception might atrophy or degenerate if stimulation is not forthcoming during an early period of language development. Eimas states that "The course of development of phonetic competence is one characterized by a loss of abilities over time if specific experience is not forthcoming," p. 346. Thus, like the adult, if phonetic differences are not used distinctively in the language-learning environment of an infant, sensitivity to the relevant acoustic attributes of these speech sounds may be attenuated and the child may fail to develop the specific mechanisms needed to discriminate the differences between these sounds (see also Werker, 1989). Eimas (1978) has argued further that the lack of experience with particular phonetic contrasts in the local environment during language acquisition may have the effect of modifying the appropriate *phonetic feature detectors* by reducing their sensitivity to specific acoustic cues in the speech signal. Thus, some detectors that were originally designed to process certain phonetic distinctions in speech may be "captured" or "subsumed" by other detectors after exposure to particular acoustic signals in the language learning environment. These detectors might, therefore, assume the specificity for only those attributes present in the stimuli to which they have been exposed. As a consequence, then, the poor discrimination observed for some phonetic contrasts might actually be due to the modification of low-level sensory mechanisms employed in discrimination of these acoustic attributes. If this view of development is correct, it would imply that mature adults would never be able to reacquire a phonetic contrast that was not present in their language-learning environment (see however, Best et al. 1988; Werker, 1989).

These conclusions concerning the role of linguistic experience in speech discrimination have become widely accepted in the literature on speech perception despite the existence of several studies demonstrating that subjects can discriminate small differences between speech sounds that were identified as belonging to the same phonological category (see Pisoni, 1973; Pisoni & Lazarus, 1974; Pisoni & Tash, 1974; Streeter, 1976a, b). When the experimental conditions are modified to reduce uncertainty or when the subjects' attention is explicitly directed to the acoustic differences between stimuli rather than to their phonetic qualities, subjects can accurately discriminate very small differences in VOT (see also Carney, Widin & Viemeister, 1977). These findings undermine the general conclusion prevalent in the literature for over thirty years - namely, that subjects cannot discriminate between speech sounds unless they are used distinctively in their native language. Nevertheless, the strong claim about the role of early experience continues to be made in the literature (see Strange & Dittman, 1984; Werker, 1989).

In a review of the effects of linguistic experience on speech perception, Strange and Jenkins (1978) concluded that the use of laboratory training techniques with adult subjects was generally ineffective in promoting enhanced discrimination of phonetic contrasts that were not employed phonemically in the subject's native language. After reading this chapter and examining results from the training experiments carried out by Strange (1972), we became interested in reexamining the performance of adults in identifying and discriminating

VOT contrasts that were not phonemically distinctive in their native language (see Pisoni et al., 1982). In particular, we wanted to know why previous attempts to use laboratory training procedures appeared to be so uniformly unsuccessful in producing changes in the perception of VOT. Given the previous work from our laboratory which demonstrated that five and six-month old infants from English speaking environments *could* discriminate both lead and lag contrasts from a VOT continuum (Aslin et al., 1981), we fully expected that native English-speaking adults would be able to discriminate these VOT contrasts as well, unless there was a real sensory loss in their underlying perceptual abilities.

In carrying out this training study, we were also interested in determining precisely how much training and experience would be required for adult English listeners to "reacquire" a nondistinctive perceptual category in voicing; whether it could be accomplished easily in the laboratory in just a few hours, or whether it would require substantially more experience and training to produce reliable changes in both identification and discrimination performance.

Insert Figure 3 about here

The results of our first training experiment on VOT are shown in Figure 3. Two groups of naive subjects were brought into the laboratory for two days and were required to identify a set of synthetic stimuli varying in VOT twice. In the first condition, subjects used only two response categories corresponding to the phonemes /b/ and /p/. In the second condition, they were given three response alternatives corresponding to [b], [p] and [ph]. The conditions were counter-balanced across both groups over the two-day period. As expected, subjects showed very reliable and consistent two-category identification functions for the English voicing categories. More interestingly, both groups of subjects also were able to reliably identify stimuli into a third perceptual category, a category with VOT values in the voicing lead region of the continuum that are not phonologically distinctive in English. Only two out of the twenty subjects we tested failed to use three responses at all. Although there was some variability in the labeling data for individual subjects, there was also a surprising amount of consistency among most of the subjects as shown in these group data.

Another experiment was also carried out with two additional groups of subjects using the same stimuli, procedures, and methodology. However, now subjects were required to identify and discriminate the same synthetic stimuli twice, once using two response categories and once using three response categories. The average identification and ABX discrimination functions from this experiment are shown in Figure 4.

Insert Figure 4 about here

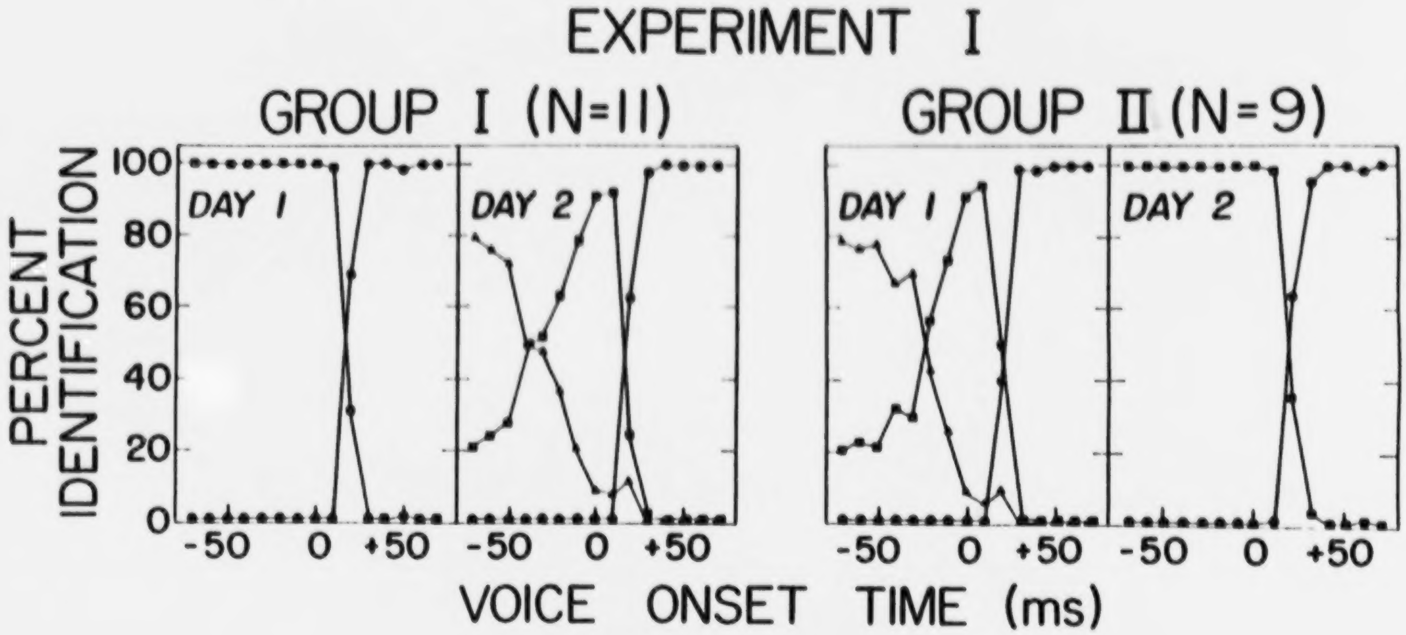


Figure 3. Average identification functions for two-and three-category labeling of synthetic speech stimuli differing in voice onset time. (From Pisoni et al., 1982).

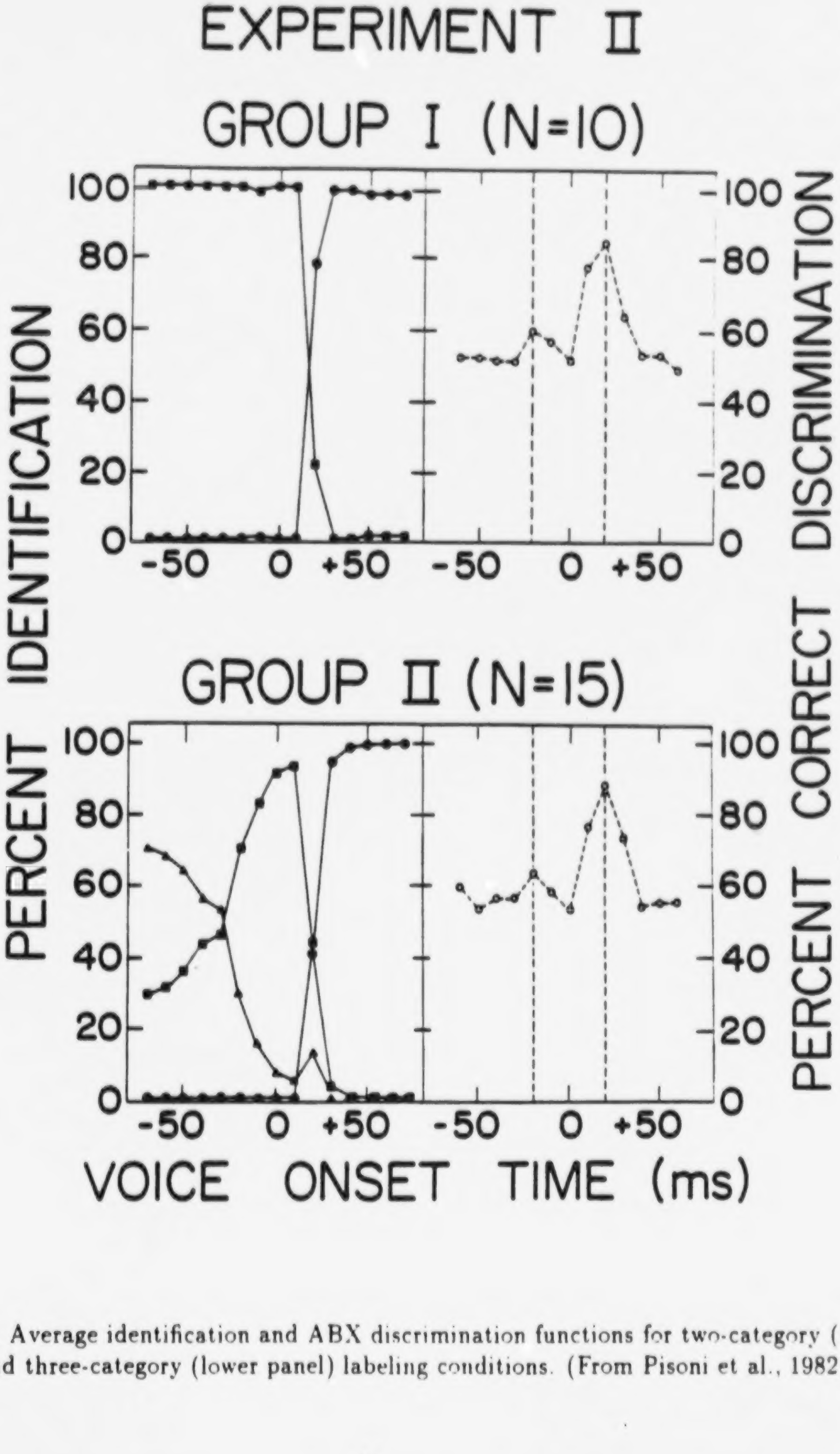


Figure 4. Average identification and ABX discrimination functions for two-category (upper panel) and three-category (lower panel) labeling conditions. (From Pisoni et al., 1982)

The two- and three-category identification functions shown in the left-hand panel of each figure are quite similar to those obtained in the first experiment. Although the average two-category data shown here are consistent and representative of individual subjects, the average three-category data are less consistent and show greater variability in the minus region of the VOT continuum.

Examination of the average ABX discrimination functions shown in the right-hand panels of Figure 4 reveals the presence of two distinct peaks in discrimination regardless of prior labeling experience. The larger peak occurs in the voicing lag region of the continuum at roughly 20 ms whereas a smaller peak can be observed in the voicing lead region at roughly -20 ms. It should be emphasized here that the subjects in the two-category labeling condition showed evidence of discriminating stimuli in the voicing lead region of the stimulus continuum despite the fact that these stimuli were all identified as belonging to the same perceptual category. Such a finding is not surprising given previous demonstrations of within category discrimination in speech perception (Pisoni & Lazarus, 1974). However, it should be noted that no special efforts were made to control or direct the subjects' attention to the differences between stimuli in this region of the VOT continuum or to modify the discrimination task to improve subjects'sensitivity.

Despite variability among individual subjects, the results of these two experiments indicate that a large majority of naive, unselected subjects can identify and discriminate an additional perceptual category in voicing quite easily without any special training or feedback. The differences in the voicing lead region are apparently discriminable, and subjects can reliably identify these sounds if given the opportunity to do this with an additional response category. Given the strong conclusions by Strange and Jenkins (1978), about the difficulty of discriminating these differences in VOT, we were surprised with the results obtained with such simple experimental manipulations.

In order to reduce intersubject variability and increase response consistency in perceptual categorization, we carried out another experiment. To accomplish this in a relatively short period of time, we used a discrimination training procedure with immediate feedback after exposure to representative exemplars of the three voicing categories. The training sequences presented only three stimuli, one representative token of each of the three voicing types, arranged in a predictable order. After the training phase was completed, subjects who met a predetermined performance criterion in identification were selected for subsequent testing in which both identification and ABX discrimination data were collected. The purpose of this experiment, therefore, was to determine if subjects who received a brief period of training would show more robust perceptual data: that is, steeper slopes in identification and heightened peaks in ABX discrimination at both voicing boundaries.

Of the original twelve subjects we recruited, six passed the 85 percent criterion on Day 1 and were invited back for the remaining sessions. Subjects who failed to meet this criterion all responded to the three training stimuli at levels well above chance, although they did not reach the required performance level. Since the previous experiment demonstrated some variability among individual subjects in VOT identification, these results were anticipated.

Insert Figure 5 about here

The average identification functions for the six criterion subjects are shown in the left-hand panel of Figure 5. These are the data collected on Day 2 of testing. As expected, these six subjects showed a high level of consistency in labeling stimuli in the voicing lead region of the continuum despite receiving only a very modest number of training trails on the three VOT exemplars (-70, 0, +70 ms). Moreover, the very steep slopes in the group identification function indicates the presence of three discrete and well-defined perceptual categories. The slope in the minus VOT region is much steeper in this experiment than in the previous experiments in which no specific training procedures were used.

The average identification and ABX discrimination data collected on Days 3 and 4 are shown in the right-hand panel of Figure 5. As observed in the previous experiment, the ABX discrimination functions obtained here also show peaks, corresponding to the boundaries between the voicing categories, and troughs, corresponding to the centers of well-defined perceptual categories. The results of this training study also demonstrate that native English-speaking adults can reacquire non-native contrasts in voicing and they can accomplish this relatively easily in a short period of time using simple laboratory training techniques.

Insert Figure 6 about here

In another training study from our laboratory, McClasky, Pisoni, and Carrell (1983) showed that knowledge about VOT perception gained from discrimination training on one place of articulation (e.g., labial) can be transferred readily to another place of articulation (e.g., alveolar) without any additional training on the specific test stimuli. The results of the transfer experiment for one group of subjects are shown in Figure 6. Apparently, naive subjects can learn very detailed and specific information about the temporal and spectral properties of VOT that is independent of the specific stimuli used in the original training sessions.

Taken together, the results of our training experiments on voicing perception demonstrate quite clearly that naive English listeners can reliably perceive differences in the minus region of the VOT continuum. The findings differ markedly from the results reported in earlier investigations of VOT perception by Strange and Jenkins (1978) which indicated that prior

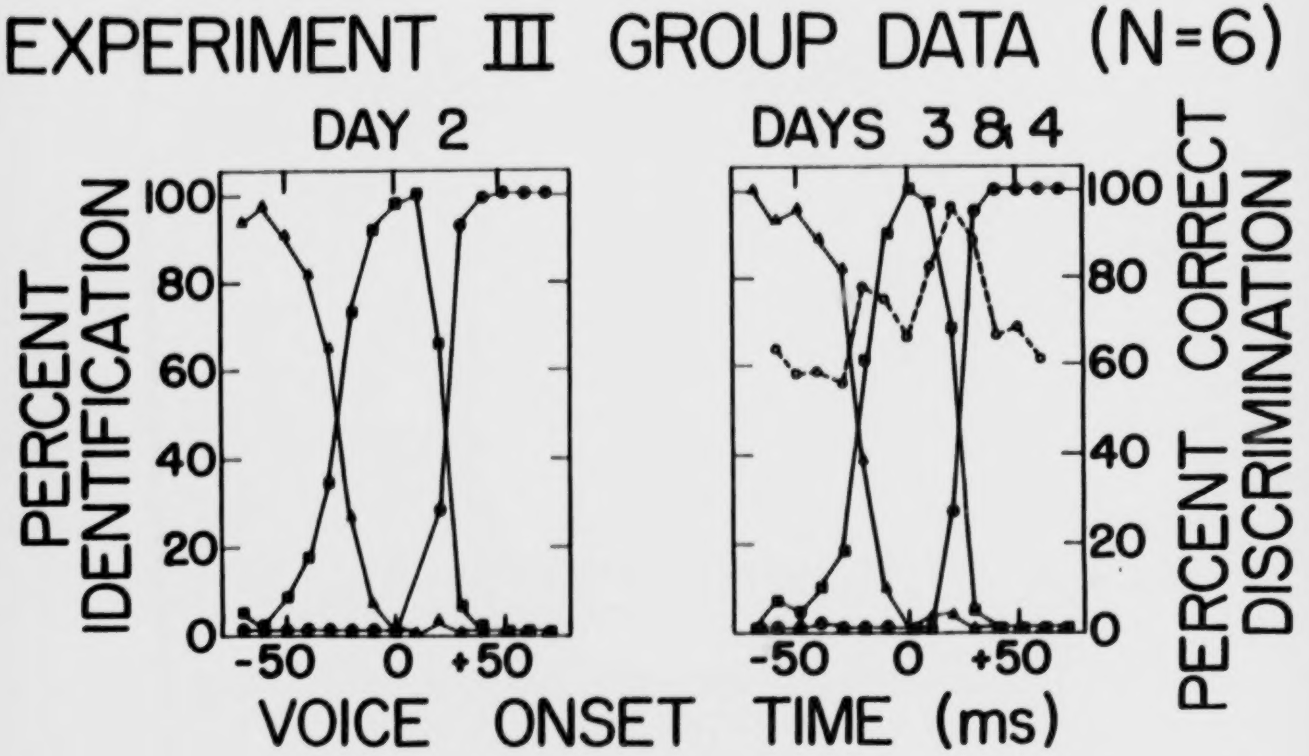


Figure 5. Average identification and ABX discrimination functions for subjects meeting an 85% identification criterion. Data in the left panel show the average identification function on Day 2; the data in the right panel show both average identification and ABX discrimination functions combined over Days 3 and 4. (From Pisoni et al., 1982).

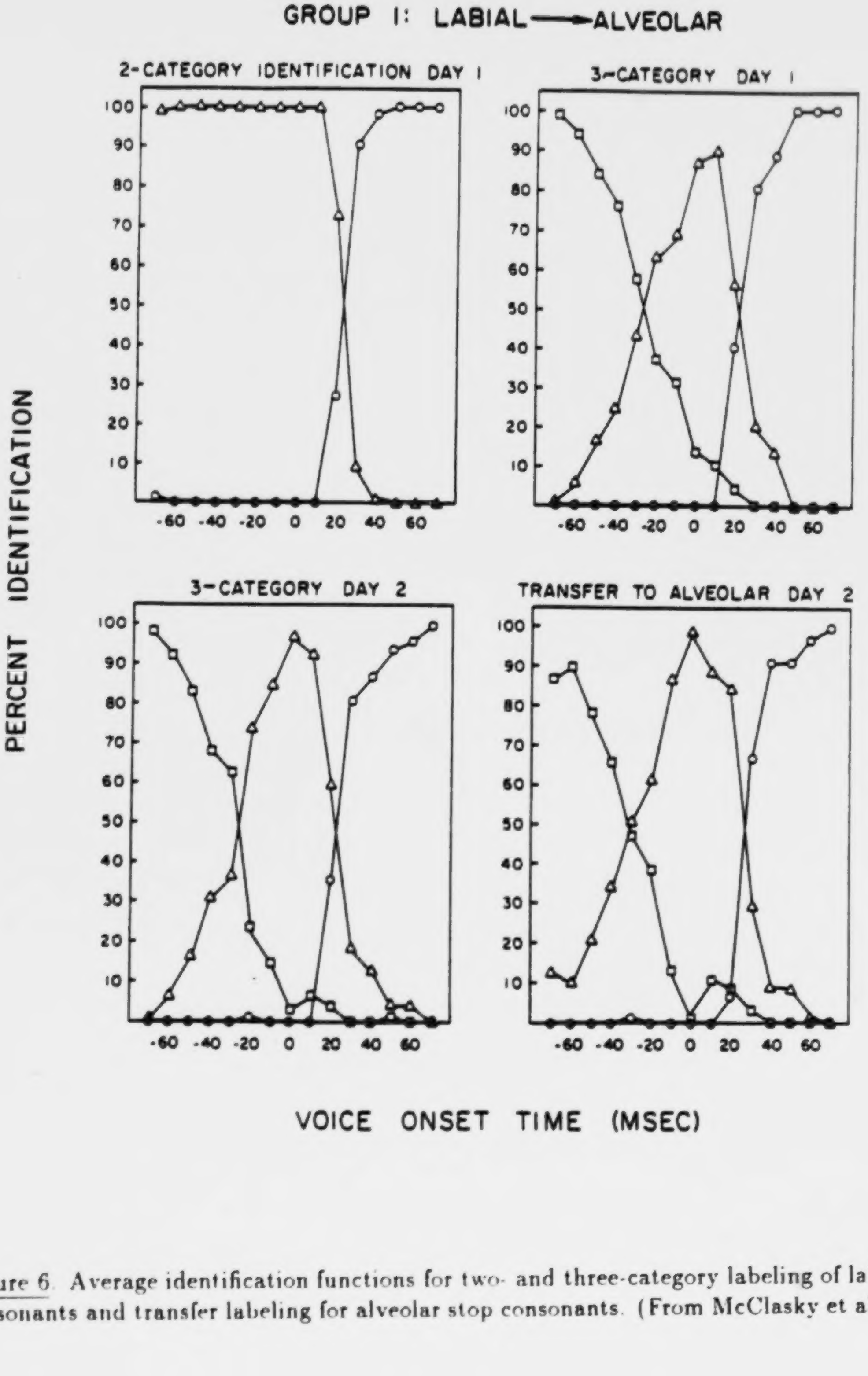


Figure 6. Average identification functions for two- and three-category labeling of labial stop consonants and transfer labeling for alveolar stop consonants. (From McClasky et al., 1983)

linguistic experience substantially diminishes perceptual sensitivity to nonphonemic voicing contrasts in adults. Our results also contradict the major conclusions of Strange and Jenkins (1978) that short-term laboratory training procedures are ineffective in modifying speech perception (see also Strange & Dittman, 1984). Given appropriate experimental procedures, our results show that naive subjects can quite easily perceive an additional perceptual contrast in voicing in the laboratory after a very short training period and they can transfer their knowledge of VOT to new stimuli with a different place of articulation that they were never trained on before. Our findings are robust and reliable and demonstrate that the underlying sensory-perceptual mechanisms have not been permanently modified or lost by prior linguistic experience.

Why have previous researchers been unsuccessful in selectively modifying the perception of VOT in adults? Is there something peculiar about the specific speech stimuli used, or might the differences be a consequence of the particular experimental methodologies employed? To answer these questions, let us turn first to an examination of the earliest cross-language speech perception experiments on VOT carried out by Lisker and Abramson (1967). They found that subjects could readily identify synthetic VOT stimuli into the phonological categories of their native language. Subjects were required to name the initial stop consonant by identifying it with one or another words in their language. Unfortunately, as far as we know, Lisker and Abramson never asked their subjects to identify the synthetic stimuli into additional perceptual categories in any of their experiments.

Although subjects in the Lisker and Abramson cross-language experiments might have been able to use additional categories by having more response choices available to them in identification, the results of oddity discrimination tests indicated that their subjects apparently could not reliably discriminate within-category differences in VOT. When discrimination is measured in the "oddity" paradigm, subjects are strongly encouraged to adopt a "context-coding" mode of response (Durlach & Braida, 1969). That is, the stimuli are immediately recoded into a more durable phonological form for maintenance in short-term memory in order to solve the discrimination problem (see Pisoni, 1973, 1975). Such a "context-coding" mode of perception is also favored by the high uncertainty conditions of the oddity discrimination task brought about by the use of a roving standard from trial to trial which effectively mixes "easy" trials with "hard" trials. Finally, immediate feedback was not provided during identification or discrimination testing. The absence of feedback in complex discrimination tasks like the oddity procudure promotes the use of highly overlearned, familiar phoneme labels and discourages fine discrimination of phonologically nondistinctive information.

Under testing conditions such as these, naive listeners apparently have great difficulty in determining precisely which acoustic attributes of the speech signal they are supposed to attend to, and which ones they are to ignore. Thus, subjects may consistently fail to discriminate fine phonetic differences within a perceptual category if they adopt a very lax criterion for detecting small differences between speech sounds. Taken together, the present

29

results demonstrate that the poor performance in VOT discrimination in earlier studies clearly is not due to a capacity limitation of any kind in processing the sensory input. We suspect that the particular combination of experimental tasks and their order of presentation to subjects may have been the major methodological factors responsible for the observed relations between identification and discrimination found by Lisker and Abramson in their well-known cross-language investigations of voicing.

In an experiment specifically designed to study the learning of a new contrast in voicing, Lisker (1970) attempted to train native speakers of Russian to distinguish between voiceless unaspirated and voiceless aspirated stops, a voicing contrast that is distinctive in English but not in Russian. In this task, although the Russian subjects learned to identify the endpoint stimuli (i.e., +10 and +60 ms VOT) slightly better than chance, their performance was not the same for both stimuli. While the majority of Lisker's subjects could differentiate the training stimuli and apparently could use two discrete labeling responses, their performance on this task was not always very consistent or reliable. Since immediate feedback for correct responses in identification was also not provided after each training trail, the subjects probably had a great deal of difficulty in determining what specific acoustic attributes of the stimuli they were to attend to selectively.

Another attempt to modify voicing perception in adults was carried out by Strange (1972) who tried to train a small number of college-age students to identify and discriminate differences in VOT in the lead region of the continuum where the Thai voiced/voiceless unaspirate boundary occurs. In her first study, four subjects received training in the oddity discrimination paradigm with "right"-"wrong" feedback provided verbally by the experimenter after each trial. When the training phase was completed, subjects carried out the oddity discrimination task without feedback. In comparison to the pretest data, all four subjects showed slightly improved overall oddity discrimination performance on the VOT stimuli during the posttest. However, no improvement was observed for discrimination of pairs of stimuli straddling the Thai labeling boundary at -20 ms VOT. The greatest increase in discrimination occurred for stimuli adjacent to the voicing boundary in English. Based on these results, Strange (1972) concluded that her subjects did not "learn" to discriminate the VOT dimension as native Thai-speaking subjects typically do. Moreover, she concluded that "there is no prepotency for adult native English speakers to discriminate differences in the region of the Thai prevoiced-voiced boundary that can be easily realized by mere practice with feedback" (p. 40).

In the second study, Strange (1972) trained three subjects to identify the members of a truncated apical series of VOT stimuli (i.e., -100 to +10 ms) into two perceptual categories. Initial training involved presentation of the endpoint stimuli in alternation without immediate feedback. However, subjects were told the number of errors they made after each block of trails by the experimenter. Oddity discrimination testing was carried out after labeling and the results showed some evidence for a slight increase in discrimination at the boundary between these two new perceptual categories. However, identification and discrimination tests using a labial VOT series failed to show any transfer of training from one VOT series to another. Nevertheless, subjects in this experiment were able to reliably identify members of the truncated apical-place series into two categories and, moreover, this labeling experience was carried over to discrimination of the same series.

Strange (1972) also carried out a third training study using a scaling procedure. Subjects were required to rate each stimulus along a scale between two endpoint reference stimuli. This procedure was adopted as a way of training subjects to perceive the VOT dimension as an acoustic continuum rather than directing their attention to discrete labeling responses as in the identification task. After training in the scaling task, subjects also carried out oddity discrimination. Although the results of this study were complicated by very high subject variability in both tasks, there was some weak evidence that training with the scaling procedure did produce effects on pereption of VOT. Posttest results for some subjects showed a shift in the scaling responses toward more gradual or continuous functions. The oddity discrimination results were more inconsistent. Some subjects showed an overall improvement in discrimination whereas others did not. As in Strange's second experiment, no consistent transfer effects from one VOT series to another were observed. Based on the outcome of these three training experiments, Strange and Jenkins (1978) offered the following summary conclusions about the effects of laboratory training in speech perception:

> "The results of these three studies show that, in general, changing the perception of VOT dimensions by adult English speakers is not easily accomplished by techniques that involved several hours of practice spread over several sessions. Although performance on each of the kinds of tests did change somewhat with experience, only the identification training task (which involved practice with general feedback only) produced categorical results approaching those found for native speakers of Thai." (p. 154).

When the results of our recent experiments on VOT are considered in light of these previous findings and the conclusions of Strange and Jenkins, it is apparent that numerous methodological factors contributed to the poor performance observed by other investigators. Nevertheless, it has generally been assumed that the failure to "learn" to perceive a new voicing contrast was somehow related to a permanent change of the perceptual or sensory mechanisms of the listener. We believe that there is little solid empirical evidence that the underlying sensory or perceptual apparatus used in voicing perception has been "retuned" or modified in any permanent manner as a result of selective early experience. Our results suggest that the perceptual selectivity observed in almost all of the previous studies on VOT perception is a consequence of attentive processes brought about by exposure to a specific subset of distinctive acoustic attributes used in the phonological system of the listener's native language.

In short, mature English adults appear to be quite capable of discriminating and categorizing acoustic information that is not phonologically distinctive in their native language.

We conclude that the underlying sensory, perceptual and cognitive mechanisms are not lost or realigned and that the attentional strategies used in speech perception are far from being as rigid and finely tuned as a number of investigators have assumed in the past. These conclusions are appropriate for voicing perception in stops. However, it remains to be seen if they apply to the perception of other speech contrasts as well. In the next section we consider the case of /r/ and /l/ perception which differs in several important respects from the voicing contrast.

Perception of /r/ and /l/

A great deal of research in speech perception has been concerned with the perception of VOT in stop consonants. This was due, in part, to the availability of high-quality synthetic stimuli which could be used quite easily to test interesting experimental hypotheses in new paradigms using a variety of subject populations (Eimas, Siqueland, Jusczyk & Vigorito, 1971; Kuhl & Miller, 1975; Streeter, 1976a,b; Lasky, Syrdal-Lasky & Klein, 1975). More recently, investigators have turned their attention to a wide variety of other phonetic contrasts in order to study the effects of early experience on perceptual development (Best et al., 1988; Werker & Tees, 1984; Werker, 1989). One speech contrast that has been investigated in some detail is the /r/ versus /l/ distinction in English (Goto, 1971; Mochizuki, 1981).

In the first cross-language study of /r/ and /l/, Goto (1971) studied a group of native Japanese subjects who were fluent in English and found that they had great difficulty discriminating /r/ and /l/ produced by native English speakers even though they could produce the contrast reliably in their own utterances. Miyawaki, Strange, Verbrugge, Liberman, Jenkins, and Fujimura (1975) tested both English and Japanese listeners with a set of synthetic speech stimuli and a set of nonspeech control stimuli containing the formant transitions appropriate for /r/ and /l/. Both groups of subjects were required to discriminate pairs of stimuli selected from each test series using an oddity test. For the English listeners, discrimination of the speech stimuli was nearly categorical. That is, discrimination of pairs of stimuli that were perceived as different phonemes was very good, whereas discrimination of pairs of stimuli that were perceived as the same phoneme was very poor. In contrast, for the Japanese listeners, discrimination of the speech stimuli was close to chance for all comparisons. Discrimination of the nonspeech stimuli, on the other hand, was significantly above chance and was comparable for both the English and Japanese listeners. The results of this study were interpreted by Miyawaki and her colleagues as additional support for an effect of linguistic experience on speech perception. Familiarity with the /r/ - /l/ distinction plays a major role in a listener's ability to correctly discriminate these stimuli. Furthermore, the differences in discrimination between the speech and nonspeech stimuli suggested that the effects of linguistic experience are apparently restricted to the phonetic coding of the acoustic signals as speech and not to the sensory processing of the underlying acoustic cues to the /r/ versus /l/ contrast. The lack of any differences in discrimination of the nonspeech stimuli by the two groups of listeners demonstrated that the nonspeech stimuli were processed equivalently on an auditory basis and were not affected by prior differential linguistic experience. This finding was obtained with the same identical acoustic cue for the /r/ versus /l/ distinction, although it was presented in isolation in the nonspeech case.

These earlier findings on /r/ and /l/ are also consistent with other, more recent studies of the perception of /r/ and /l/ by Japanese listeners. MacKain, Best and Strange (1981) found that, even after several years of living in an English-speaking environment, adult Japanese listeners still differ in several ways from native speakers of English in their identification and discrimination of synthetic /r/ and /l/ stimuli. In addition, Sheldon and Strange (1982) showed that Japanese listeners even have difficulty perceiving natural tokens of /r/ and /l/ produced by native speakers of English. Other studies using Japanese listeners have shown that the perception of /r/ and /l/ is highly context dependent (Gillette, 1980; Mochizuki, 1982; Sheldon & Strange, 1982). Performance is generally lowest for perception of /r/ and /l/ in initial singleton or initial clusters and highest for /r/ and /l/ in final position. While there are no obvious phonological reasons for these context effects, acoustic analyses of /r/ and /l/ in several different phonetic environments have revealed large and systematic differences in the durations of the formant transitions (Dissosway-Huff, Port & Pisoni, 1982).

In a more recent study, Strange and Dittman (1984) attempted to modify the perception of /r/ and /l/ in Japanese listeners using several laboratory training procedures. Although Strange and Dittman (1984) were primarily concerned in their study with assessing generalization of the training procedures to naturally produced English words, they did raise several important criticisms of the earlier training studies of VOT carried out by Pisoni and his colleagues. Their criticisms of our training experiments and many of the previous studies of VOT perception are well motivated in our view and will be summarized below because they played an important role in the design of Strange and Dittman's study and in our own work reported below.

First, Strange and Dittman note that earlier training studies used highly controlled synthetic stimuli instead of tokens of natural speech. When synthetic speech stimuli are used in perceptual experiments subjects are exposed to highly improverished stimuli that contain only the minimal acoustic cues that are necessary to distinguish a particular phonetic contrast (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). In contrast, natural speech is extremely redundant. Each phonetic contrast has multiple acoustic cues encoded in the speech signal which maintain intelligibility under very adverse conditions. When synthetic speech stimuli are used in training experiments, it is very likely that listeners will focus their attention only on the cues that are present in the signal and fail to generalize to other stimuli containing multiple redundant cues to the same phonetic contrast. In this connection, it is interesting to note that Mochizuki (1981) actually found very high levels of performance for naturally produced tokens of /r/ and /l/ although only the results from her synthetic speech conditions appear to be cited in the literature (MacKain, Best & Strange, 1981; Strange & Dittman, 1984).

Second, Strange and Dittman point out that all of the previous training studies used nonsense syllables rather than real English words. The use of nonsense syllables as stimuli in training experiments is problematic for several reasons. First, nonsense syllables remove any lexical contributions to recognition and consequently focus the listener's attention on only the individual phonemes that distinguish the test syllables. Second, in most of the previous training studies that used nonsense syllables as stimuli, the range of phonetic environments was very small. Thus, subjects received very little stimulus variability during learning. The lack of stimulus variability may prevent the development of robust perceptual categories that would be helpful in later tests of generalization with real words where there is typically a great deal of variability across different phonetic environments.

Third, Strange and Dittman argue that there are important differences in the phonetic and phonological distributional properties of voicing in stop consonants compared with the distributional properties of /r/ and /l/. In particular, the voicing contrasts that were used in the previous training studies on VOT perception were allophonic in English. Listeners were, in fact, exposed to these sounds in their environment even though the contrasts are not used distinctively. In contrast, as Strange and Dittman point out, this is not true for the phonemes /r/ and /l/ which do not occur as allophones in Japanese. Thus, native speakers of Japanese are never exposed to these contrasts during language acquisition (Werker, 1989).

Finally, Strange and Dittman (1984) note that the acoustic cues underlying the voicing distinction in stops are markedly different from the complex temporal and spectral changes that are used to distinguish the phonemes /r/ and /l/ in various phonetic environments. Their argument here is that voicing may somehow be psychophysically more distinctive or "robust" and therefore much more discriminable to listeners than the acoustic cues that underlie other speech contrasts (see Burnham, 1986). Because the acoustic correlates of phonetic contrasts differ widely and, therefore, have quite different psychological spaces, it is often difficult to equate the underlying sensory scales (Lane, 1965). However, if a phonetic contrast is discriminable on a psychophysical basis (i.e., differences are above threshold) then the relative differences in perception between various speech contrasts must be considered within the domain of selective attention, rather than viewed simply as a basic limitation on sensory processing of the stimulus input (see Nosofsky, 1986, 1987). The distinction between a "true" sensory loss and a loss due to selective attention has not been widely recognized in the speech perception literature and is often treated as having the same underlying basis (see Burnham, 1986).

Given these criticisms of the earlier training studies on VOT, Strange and Dittman (1984) attempted to modify Japanese listeners' perception of /r/ and /l/ in a set of naturally produced real English words. A pretest-posttest design was used with the same set of natural speech tokens to assess the effects of discrimination training with a synthetic "rock-lock" continuum. Subjects were required to identify a member of a minimal pair using a two-alternative forced-choice identification test. The effectiveness of the training procedures was assessed by comparing the initial levels of performance with naturally produced words to performance after discrimination training with the synthetic speech.

Strange and Dittman (1984) found that, although discrimination performance improved gradually for all subjects over the training sessions with the synthetic speech series, the effects of discrimination training apparently did not generalize at all to the naturally produced real English words used in the posttest. Comparisons of pretraining and posttraining categorical perception tests using the synthetic "rock-lock" training stimuli did show some changes in performance for seven of the eight subjects. And, five of the seven subjects evidently also showed improvement and more categorical-like perception in identification and oddity discrimination tests on an acoustically dissimilar "rake-lake" synthetic test series. However, transfer of training did not generalize to identifying minimal pairs of naturally produced English words that contrasted /r/ and /l/. Based on these results, Strange and Dittman (1984) concluded that "modification of perception of some phonetic contrasts in adulthood is slow and effortful" and "required intensive instruction and considerable time and effort at least for some types of phonetic contrasts."

As in the studies of voicing perception described earlier, we believe that a number of factors may have been responsible for Strange and Dittman's failure to find improvement in the perception on /r/ and /l/ in naturally produced words after discrimination training. Some of these factors are primarily methodological in nature and are easy to modify but others are more conceptual in scope and reflect deep theoretical biases. A close examination of the design of Strange and Dittman's study reveals a number of important theoretical assumptions that were made about what listeners are actually learning in laboratory training experiments of this kind. An examination of these assumptions provides some insight into what Strange and Dittman's subjects were learning in their training experiment and why they failed to show any evidence of generalization to naturally produced English words.

First, let us consider the AX discrimination training procedure that Strange and Dittman used. Based on earlier successful work of Carney et al. (1977), this procedure was employed to improve listeners' perception of within category acoustic differences by focusing attention on the subtle acoustic cues that differentiate synthetic tokens of "rock" and "lock." There is now an extensive literature demonstrating that low-level sensory information is extremely fragile and often quite difficult to maintain in sensory memory without additional recoding into more permanent representations in short-term memory (Shiffrin, 1976). Except under special testing conditions like the ones used by Carney et al. (1977), listeners in speech perception experiments typically have access to only the product of this process, namely, the phonetic representations, and not the intermediate forms (see Pisoni, 1973).

It appears very likely to us that discrimination training procedures like the AX task that focus the listener's attention on low-level acoustic information in sensory memory will probably not be very successful in promoting generalization to more robust conditions in which naturally produced real words are used as test stimuli (see also Jamieson & Morosan, 1986). Discrimination training may generalize from the specific training stimuli to other synthetic stimuli that have contrasts in the same identical phonetic environments but it seems unlikely to us that training listeners to perceive small within category differences between synthetic /r/ and /l/ will be of much help in identifying these phonetic contrasts in other environments using natural speech where there is typically a great deal of acoustic-phonetic variability. The outcome of Strange and Dittman's study is therefore not at all surprising to us and is entirely consistent with this explanation of the extent to which this particular kind of discrimination training will generalize to novel tokens of /r/ and /l/ in natural speech. Jamieson and Morosan (1986) have made similar points with regard to designing training methods to modify phonetic perception.

A second issue concerns the theoretical assumptions surrounding what listeners are actually learning and what kind of knowledge they are acquiring in discrimination training experiments such as these. While not made explicit anywhere in their paper, Strange and Dittman apparently assumed that by training Japanese subjects on /r/ and /l/ in initial syllable position, their subjects would somehow be able to generalize what they learned about /r/ and /l/ to other phonetic environments that were not explicitly presented during training. We believe this assumption implies that subjects are learning about fairly abstract perceptual units such as phonemes in discrimination training and that the perceptual learning that goes on is context-independent.

It may very well be the case, however, that during discrimination training subjects are actually acquiring highly stimulus-specific information about the acoustic cues for /r/ and /l/ in different phonetic environments that they are exposed to and that the training and knowledge gained from one phonetic environment may not generalize to other environments without explicit presentation of exemplars from these environments. Again, the results reported by Strange and Dittman are consistent with this observation. They found some improvements in identification and discrimination of synthetic tokens that were phonetically similar to the stimuli used in training, but they failed to find any evidence of generalization of the training to /r/ and /l/ contrasts in new phonetic environments or in naturally produced English words. Thus, subjects were probably not learning about abstract context-independent perceptual units such as phonemes but, instead, were encoding specific details of the context into their representations.

Finally, Strange and Dittman used highly controlled synthetic speech stimuli in training and in subsequent tests of identification and oddity discrimination, but they tested their subjects for generalization of /r/ and /l/ with naturally produced words using a minimal pair forced-choice identification test. In the AX discrimination training tests and the subsequent categorical perception tests, subjects are required to focus their attention on the acoustic cues used to distinguish phonemes in the same synthetic stimuli and the same phonetic environments. In contrast, in the minimal pair test, subjects are required to identify real English words contrasting in /r/ and /l/, not individual phonemes, and they are required to do this for /r/ and /l/ in a variety of new phonetic environments.

Training subjects to discriminate small acoustic differences between stimuli in a highly restricted phonetic environment may not provide subjects with robust and useful knowledge about the range of stimulus variability that can be generalized to an entirely new task – namely, the identification of English words in which the same phonemes now occur in different phonetic environments. The changes in perception, gained by discrimination training using one set of tasks, such as phoneme identification and oddity discrimination, to study categorical perception phenomena, may not be very helpful to subjects when they are required to carry out lexical analysis of an entire word using a minimal pair test. The acoustic information that subjects are trained to attend to and subsequently encode in the AX task may be useful in phomeme identification and oddity discrimination tests which require subjects to make fine "within category" discriminations. However, this kind of information may not be very helpful in discriminating phonemes that appear in different phonetic environments in naturally produced words (see also Jamieson & Morosan, 1986).

New Data on the Perception of /r/ and /l/

Recently, Logan, Lively and Pisoni (1988) carried out a training study to investigate the conditions under which a group of native Japanese speakers could learn to identify naturally produced words contrasting in /r/ and /l/ in a variety of phonetic environments. The experiment was motivated, in part, by the results of the earlier study carried out by Strange and Dittman (1984) on /r/ and /l/ and our previous training studies on the perception of VOT in stop consonants. In designing this study, we wanted to develop a set of training procedures that would not only produce changes in the perception of /r/ and /l/ in real English words, but would also prove useful in settings outside the laboratory. We began by adopting the same pretest–posttest design that Strange and Dittman (1984) used. In fact, so that direct comparisons could be made between the two studies, we used the same identical 16 minimal pairs of test words contrasting in /r/ and /l/ and the same two-alternative forced-choice identification test. However, our training procedures differed in several important ways from the methods originally used by Strange and Dittman (1984).

The first change in the training procedure involved replacing the AX discrimination test with a two-alternative identification test. This was done so that the responses used in training would be directly compatible with the responses used in generalization testing. Maintaining response compatibility throughout the experiment encouraged subjects to use the same acoustic information they attended to and encoded during training in the subsequent generalization tests. Thus, in contrast to the procedures used by Strange and Dittman, we began by having subjects identify minimal pairs of words rather than focus their attention on small within category differences between phonemes.

The second change in training involved the use of naturally produced tokens of real English words contrasting in /r/ and /l/ in five different phonetic environments. Strange and

Dittman used only synthetic speech stimuli in training and their "rock-lock" continuum differed only in syllable initial position although they tested for generalization in four different phonetic environments using naturally produced words. Again, by training subjects to selectively attend to words containing /r/ and /l/ in several different phonetic environments, we hoped that they would focus their attention and encode the relevant criterial acoustic features of these different contexts and then use this information when presented with novel words in subsequent generalization tests. Examples of the test words used in the five phonetic environments are shown in Table 1.

Insert Table 1 about here

Third, the naturally produced tokens used in training were produced by five different talkers in order to present listeners with a wide range of stimulus variability in learning. However, none of the items used during the training phase ever appeared in the pretest or posttest conditions. The pretest and posttest items were produced by different talkers than the ones used to produce the training items. This was done in order to dissociate talker-specific and item-specific learning effects. The use of multiple test items produced by several different talkers was motivated, in part, by the desire to present the subjects with a great deal of stimulus variability during training. We hoped this would encourage subjects to form "robust" phonetic representations for /r/ and /l/. Strange and Dittman trained subjects to discriminate /r/ and /l/ in only one phonetic environment using only one talker (i.e., a speech synthesizer) so it is not surprising that they failed to find any transfer between pretest and posttest performance.

Finally, in addition to assessing the effectiveness of the training procedures by measuring transfer from the pretest to the posttest, we also included two additional generalization tests with novel words contrasting in /r/ and /l/. These words were never presented either in the pretest–posttest or training phases of the experiment. One generalization test used a novel talker; whereas a second generalization test used a familiar talker who had produced a set of the training items. Both talkers produced novel words that contrasted in /r/ and /l/ in a variety of phonetic environments. We hoped these additional generalization tests would provide detailed information about what aspects of the stimuli subjects were encoding into long-term memory.

For ease of exposition, the results of our experiment will be presented in three sections below, corresponding to the pretest–posttest data, the training data and the generalization data. In all cases, we will be looking at performance in perceiving naturally produced English words containing /r/ and /l/ using the minimal pair identification test. In this procedure, subjects are required to identify a word on each trial using one of two possible response alternatives. Our subjects were six native speakers of Japanese who were enrolled as

Table 1

Phonetic Environments

(from Logan, Lively, & Pisoni, 1988)

Environments:	Examples:
1. Initial Consonant Cluster (c r/l v...)	brush-blush, grass-glass
2. Initial Singleton (r/l v c)	rake-lake, rock-lock
3. Intervocalic (... v r/l v ...)	pirate-pilot, oreo-oleo
4. Final Consonant Cluster (c v r/l c)	mart-malt, board-bold
5. Final Singleton (...v r/l)	mare-mail, pear-pail

students at Indiana University. They lived in the U.S. for periods ranging from six months to three years at the time of testing. These subjects were comparable to those used by Strange and Dittman (1984).

Insert Figure 7 about here

Pretest-Posttest Performance

The percentage of correct identification responses averaged over the six subjects in the pretest and posttest conditions is shown in Figure 7. Overall, a significant increase in performance was observed, $p < .005$. In all cases, subjects improved in their ability to identify English words containing /r/ and /l/ after the training phase. While the absolute difference of eight percent was not large, the effect is highly significant and was observed for every one of the six subjects. The present findings are quite different from those reported by Strange and Dittman using the same pretest–posttest items. They found no differences in performance between pretest and posttest after training. Our results demonstrate that the major factor distinguishing the two studies must be the training procedures since the pretest–posttest items and the minimal pair testing procedures were identical. Before examining the training data, however, it is useful to look at the pretest–posttest results broken down by phonetic environment in order to gain some insight into the nature of the perceptual learning that went on during the training phase.

Insert Figure 8 about here

The percentage of correct responses for each of the four phonetic environments in the pretest–posttest is plotted in Figure 8. The figure shows two important trends. First, overall performance on /r/ and /l/ differs substantially across the four phonetic environments. Performance is best for /r/ and /l/ as singletons in final position and worst for /r/ and /l/ in clusters in initial position. Second, the effects of training appear to have the largest influence on /r/ and /l/ in initial clusters and in intervocalic position. Training produced almost no change in performance for /r/ and /l/ as singletons in initial position or in final position. The latter is probably due to a ceiling effect because performance before training on these items was quite high. The absence of any change in performance for the singletons in initial position may be due to a variety of factors including the inherent discriminability

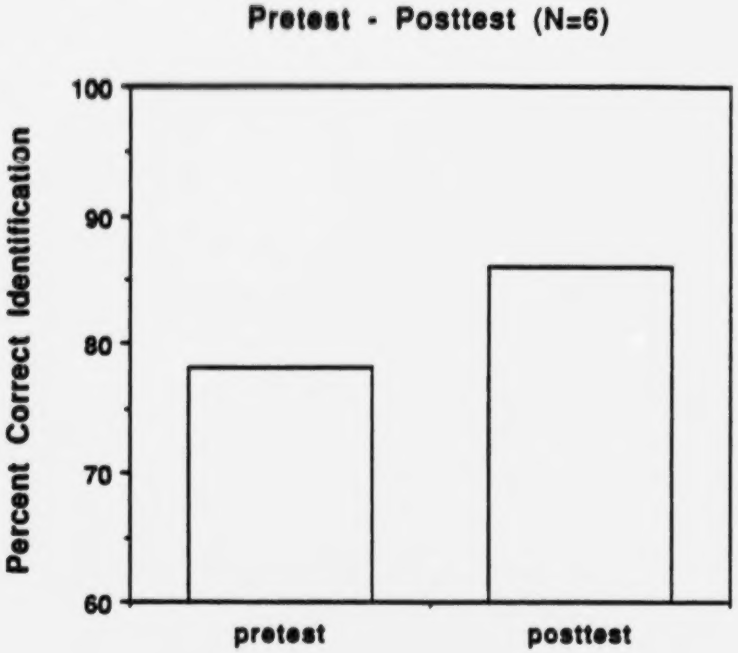


Figure 7. Average percent correct identification of test words containing /r/ and /l/ in the pretest and posttest conditions. (From Logan, Lively & Pisoni, 1988).

41

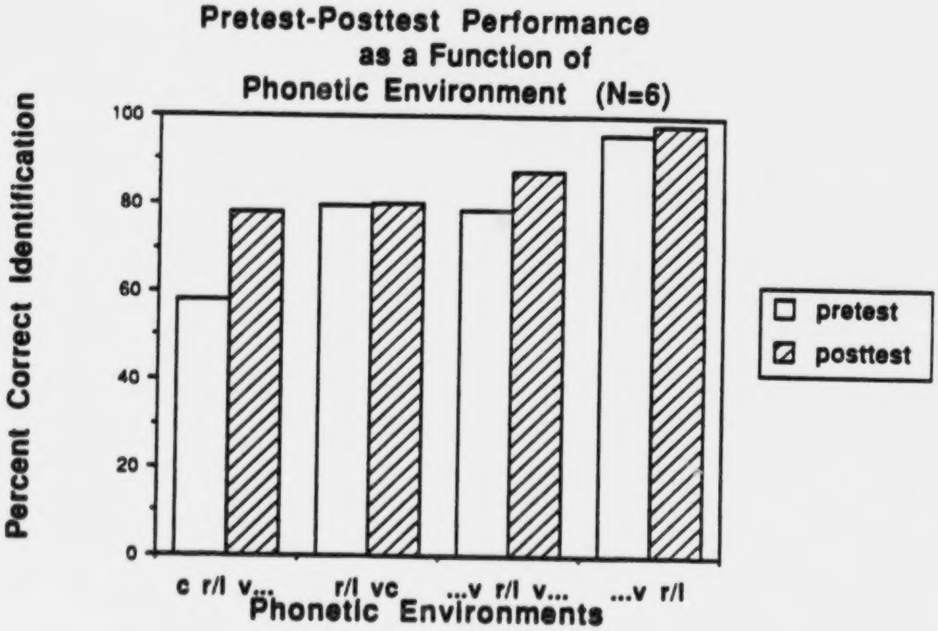


Figure 8. Average percent correct identification of test words in the pretest and posttest conditions as a function of phonetic environment. (From Logan, Lively & Piosni, 1988).

of /r/ and /l/ in this environment and the attentional focus of the listener given several potentially more salient cues to /r/ and /l/ in other phonetic environments.

Taken together, however, the present results show not only that perceptual learning took place, but that the learning was apparently highly context-dependent in nature. The non-uniformity across the four phonetic environments implies that subjects were attending to and encoding different acoustic cues for different environments they were exposed to during training. Rather than learning about an abstract context-independent unit such as a phoneme, subjects were apparently learning very detailed context-dependent information for /r/ and /l/ in these different environments.

Insert Figure 9 about here

Training

The percentage of correct responses for the training items, broken down by week, is shown in Figure 9. Performance improved over the three week training period although the largest change occured during the first two weeks. The change in performance from week 1 to week 2 was significant, but the change from week 2 to week 3 was not. Whatever subjects are learning about the cues to /r/ and /l/ in these different phonetic environments, they are apparently learning it in a relatively short period of time.

Insert Figure 10 about here

Figure 10 shows the percentage of correct responses to the training items as a function of phonetic environment. As in the pretest posttest data, performance varied over a wide range. The best performance was observed again for /r/ and /l/ as singletons in final position; the worst performance was observed for /r/ and /l/ in initial clusters. Thus, even during training with feedback, not all phonetic environments are learned equally well under these training conditions.

Insert Figure 11 about here

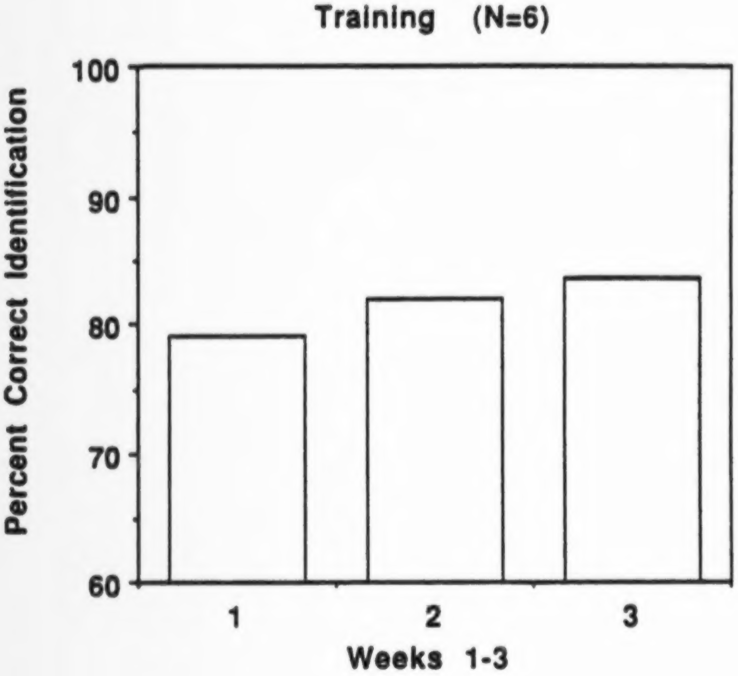


Figure 9. Average percent correct identification of words used during training as a function of week. (From Logan, Lively & Pisoni, 1988)

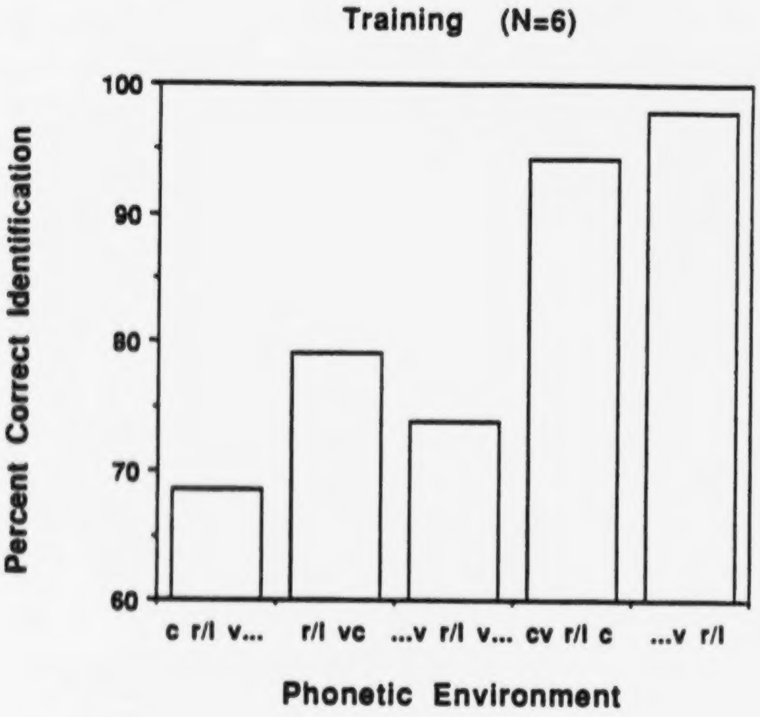


Figure 10. Average percent correct identification of words used in training as a function of phonetic environment. (From Logan, Lively & Pisoni, 1988).

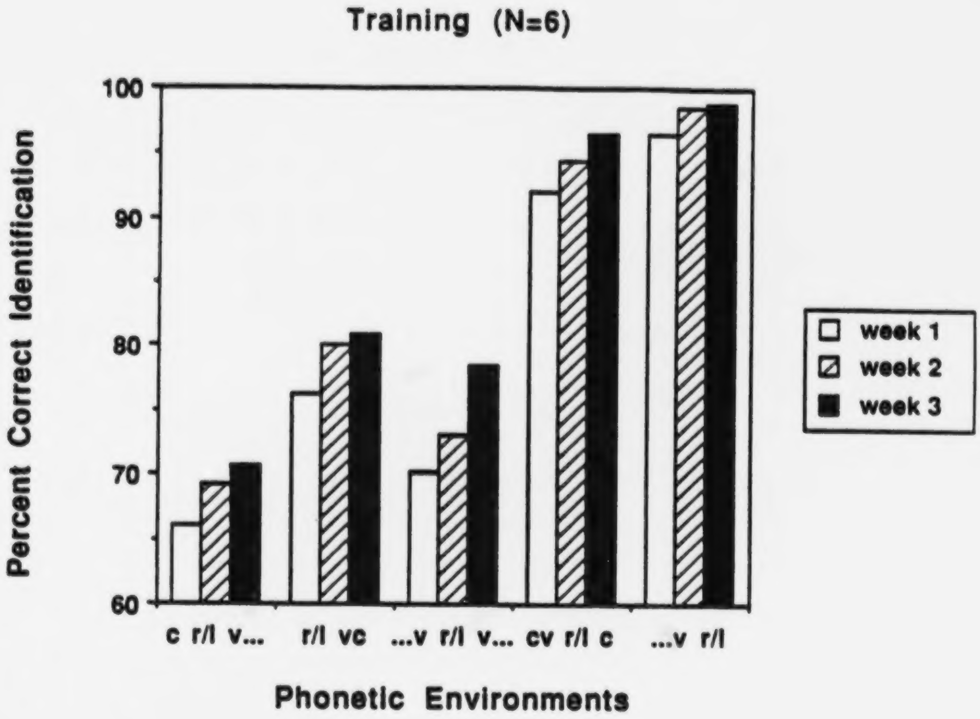


Figure 11. Average percent correct identification of words used in training as a function of week for each phonetic environment. (From Logan, Lively & Pisoni, 1988).

Figure 11 shows the effects of phonetic environment as a function of week of training. The same overall pattern of results observed for the five different phonetic environments is replicated each week suggesting the existence of inherent differences in perception of /r/ and /l/ in different phonetic environments, a finding that has been reported previously by other researchers (Mochizuki, 1981; Dissosway et al., 1982).

Insert Figure 12 about here

Because the training procedures were automated and under the control of a laboratory computer, we were able to record response times from subjects during the two-alternative identification task. Figure 12 shows the average response times for correct responses to test items in the five phonetic environments as a function of week of training. The overall pattern of the response times parallels the data shown in the previous figure for percent correct performance. Response times are fastest for /r/ and /l/ in final position and slowest for /r/ and /l/ in initial clusters and intervocalic position. For those phonetic environments in which identification was very high at the outset of training (i.e, final singletons and final clusters), the response times appear to decrease consistently each successive week. In contrast, for the other three phonetic environments, in which performance was low at the outset of training, the response times show a very different pattern over the three week period, perhaps reflecting subjects' initial difficulty in focusing their attention on the appropriate acoustic cues for /r/ and /l/ in these particular phonetic environments. The inverted U-shaped functions for these phonetic environments suggest that once subjects learned what aspects of the stimulus to attend to, their performance began to improve and their latencies decreased substantially.

Insert Figure 13 about here

In addition to variability in identification performance as a function of the phonetic environment, we also observed differences among the talkers who produced the items used in training. Figure 13 shows the percentage of correct responses in training for each of the five talkers. The test words produced by Talkers 4 and 5 were more intelligible than the words produced by the other three talkers used during training.[1] In pretesting these items with native speakers of English, no reliable differences were observed among any of the talkers.

---

[1] The order of presentation of the talkers was the same for all of the subjects, so this finding may reflect an effect of training (which in itself is interesting) rather than differences in talker intelligibility.

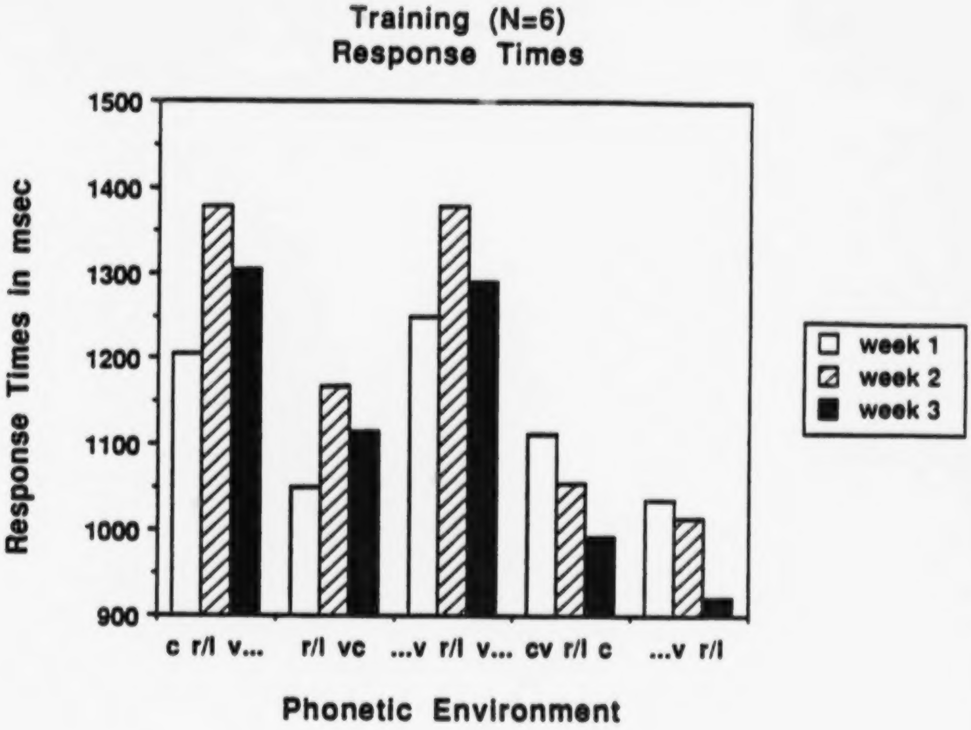


Figure 12. Average response times for correct responses to test words used in training as a function of week for each phonetic environment. (From Logan, Lively & Pisoni, 1988).

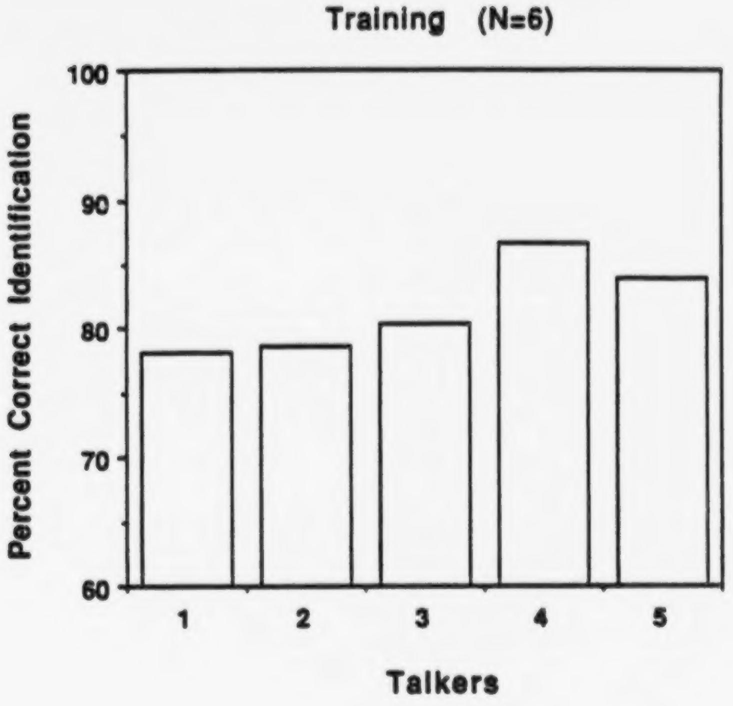


Figure 13. Average percent correct identification of words used in training as a function of talkers. (From Logan, Lively & Pisoni, 1988).

The results of the training phase of the experiment show clearly that the perceptual learning of /r/ and /l/ is highly context dependent. Large differences in accuracy and response latency were observed in identification of words containing /r/ and /l/ as a function of phonetic environment and talkers. While there was a general trend for the learning to improve over the three weeks of the experiment, the largest gains were made during the first two weeks. The pattern of response times suggests that subjects were actively trying to learn the specific criterial properties that distinguish /r/ and /l/ in different phonetic environments. Some environments appear to be relatively easy to learn whereas others appear more difficult. Indeed, even after three weeks of training during which there were large improvements in performance overall, the identification of /r/ and /l/ as singletons in initial position, did not change reliably from pretest to posttest. In contrast, perception of /r/ and /l/ in two of the other environments showed substantial changes in performance after training.

Insert Figure 14 about here

Generalization with Novel Words

Figure 14 shows the results of the two generalization tests with novel words. Condition TG1 consisted of novel words produced by a novel talker; condition TG2 consisted of novel words produced by a familiar talker (Talker 4) who was used during the training phase. These results, which are based on only three subjects, show that identification of novel words from condition TG2, a familiar talker, is better than identification of novel words from TG1, a novel talker ($p < .09$). Apparently, familiarity with a talker's voice improves the identification performance on novel words that a listener has never heard before in the experiment. Similar findings have been reported recently by Mullennix, Pisoni, and Martin, (1989).

The present results suggest the intriguing possibility that listeners are not only encoding context-sensitive information about the specific phonetic environments that /r/ and /l/ appear in during training, but they are also encoding, along with this, quite detailed acoustic-phonetic information about the specific properties of the talker's articulation as well. Acoustic information about a talker's voice may therefore be encoded along with a phonetic representation of the input and stored in long-term memory for later use (see Martin, Mullennix, Pisoni & Summers, 1989).

Stimulus variability generated by exposure to speech produced by different talkers may play a central role in developing robust phonetic categories in perception that are not defined

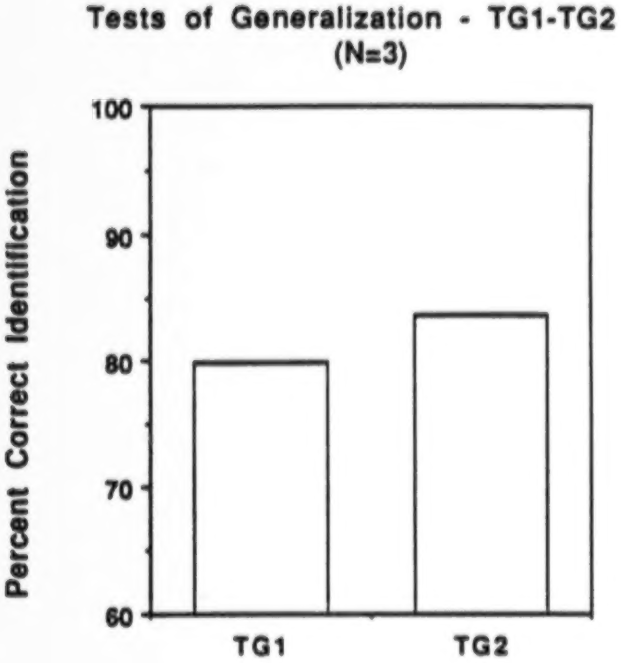


Figure 14. Average percent correct identification of novel words in generalization tests for a new talker (TG1) and a familiar talker who was used in training (TG2). (From Logan, Lively & Pisoni, 1988).

exclusively by a small number of absolute criterial acoustic features which must be present in the signal for it to be identified reliably in different environments. Stimulus variability produced through these experimental manipulations may be similar to the kind of stimulus variability encountered by non-native speakers who have had intensive conversational experience with English (MacKain, Best & Strange, 1981). Subjects involved in conversational instruction with native speakers apparently do show improved perceptual skills with non-native speech contrasts (Gillette, 1980; Mochizuki, 1981; MacKain et al., 1981). Conversational experience may therefore provide exposure to speech produced by diverse talkers producing phonetic contrasts in a wider variety of environments than would be encountered in laboratory settings. Regardless of the precise explanation, the present findings demonstrate that stimulus variability is useful in perceptual learning and may contribute to the development of more robust context-sensitive perceptual categories.

General Discussion

The results of the training experiments summarized in this chapter demonstrate quite convincingly that non-native speakers can, in fact, learn to perceive speech contrasts that are not distinctive in their native language. These findings, which were obtained with two very different phonetic contrasts - the voicing distinction in stop consonants cued by VOT and the /r/ and /l/ contrast - show that the developmental decline in discriminative capacities and associated perceptual loss is not permanent and can be "reacquired" in a relatively short period of time using relatively simple laboratory training procedures. Based on these findings, we believe that the previous conclusions about the effects of early linguistic experience on speech perception are unjustified and have been greatly exaggerated in the literature on perceptual development.

There can be little question in anyone's mind that a major aspect of the development of speech perception in infants and young children involves some form of developmental change and perceptual reorganization as a function of specific experiences in the language-learning environment. The acquisition of language, like the development of bird-song, requires an intensive period of vocal learning during which the young child begins to acquire the local dialect and lexicon of the speech produced in his/her language-learning environment. At the present time, we are just beginning to develop adequate theoretical accounts of precisely how this process takes place and how the sensory prerequisites and phonetically relevant capabilities are shaped, modified or tuned to the important phonetic distinctions in the language-learning environment (see Aslin, 1981; Aslin, 1985; Aslin & Pisoni, 1980; Studdert-Kennedy, 1986, 1987; Jusczyk, 1985, 1986).

When one considers the data from a wide variety of studies on infant speech perception that have been carried out over the last 18 years, it becomes obvious that prelinguistic infants display evidence of a universal sensitivity to phonetically relevant sound contrasts in language (for reviews see Aslin, Pisoni & Jusczyk, 1983; Jusczyk, 1985, 1986; Kuhl, 1987; Werker, 1989). These findings have demonstrated consistently that infants have the sensory and perceptual prerequisites to eventually acquire the phonetics and phonology of any spoken language. Unfortunately, relatively few of these infant studies have addressed the somewhat broader issues of the nature of the developmental change that takes place when the child begins to acquire the first rudiments of spoken language and reliably starts to assign meanings and communicative intent to sound patterns in his/her environment. And, even fewer studies have addressed the issue of the apparent developmental decline in the perceptual abilities of mature adults after they have acquired their native language (see however Aslin & Pisoni, 1980; Best et al., 1988; Flege, 1987; Walley, Pisoni & Aslin, 1981).

Some attempts have been made recently by Werker (1989) to deal with developmental change in speech perception, but her theoretical efforts to date have not been concerned with the precise mechanisms that underlie change and perceptual reorganization in speech perception. In addition, she has focused most of her work on the period immediately before the child's first words appear, thus effectively preventing her from addressing issues related to the contribution of phonology and lexical knowledge to speech perception.

Taking a different approach, Jusczyk (1985, 1986) has recently suggested that young infants actively adjust perceptual weights to encode phonologically distinctive information in their enviroment. His proposal, which we will return to in the next section, places a great deal of emphasis on attentional mechanisms in perceptual development. It therefore provides a unified way of dealing with both the infant data, showing universal phonetic sensitivity, as well as the adult data, showing perceptual reorganization and developmental decline in phonetic discrimination. However, like Werker, Jusczyk has also focused most of his efforts on the period immediately before the child's first words.

Both Ferguson (1986) and Studdert-Kennedy (1987) have remarked recently about the serious gap in our knowledge and understanding of the relations between the infant speech-sound perception data and the child phonology literature. Not only have there been few efforts to relate findings from these two areas, but there has been little useful theoretical work on the nature of the perceptual reorganization that takes place when the child begins to use spoken language in a communicatively relevant way. Previous accounts of developmental change, such as the one suggested recently by Werker (1989), have focused on a simple dichotomy between language-based and sensory-based processes. But this is obviously not sufficient to account for the present set of findings with mature adults who are able to reacquire the perceptual abilities needed to discriminate and identify phonetic contrasts that were not distinctive in their language-learning environment.

If mature adults have the basic underlying sensory abilities needed to discriminate phonetically relevant speech contrasts, as the present findings demonstrate, then why do they apparently have such great difficulty using these abilities when they are called upon to perform tasks that require linguistically relevant perceptual responses? Is it desirable to have two accounts of perceptual change: one for infants and young children acquiring their first language, and another for adults reacquiring new phonetic contrasts in a second language? Or should we attempt to develop a common approach that is appropriate for both sets of findings? As we pointed out in the introduction, contemporary theories of speech perception were formulated to deal with the mature adult and little, if any, attention has been devoted to issues of development, developmental change, or second language acquisition. In the next section, we suggest an approach that can be applied to both the adult and infant perceptual findings. Our basic argument is that these seemingly diverse findings reflect the operation of attentional processes in speech perception that are primarily perceptual rather than sensory-based in nature.

The Role of Selective Attention in Speech Perception

Although cognitive psychologists have studied selective attention for many years, it has only been recently that researchers working in speech perception have acknowledged the fundamental role of attentional processes in identification, categorization, and discrimination of speech sounds (for a review see Nusbaum & Schwab, 1986; see also Jusczyk, this volume). In a number of important studies on the identification and categorization of complex multidimensional visual stimuli, Nosofsky (1986, 1987) has shown how selective attention to specific stimulus dimensions can modify the underlying psychological space and therefore affect the perceived similarity relations among component dimensions in different tasks. According to Nosofsky, selective attention can be thought of in terms of a metaphor that involves stretching of psychological distances to attended dimensions and shrinking of distances for unattended dimensions. When subjects are required to attend selectively to one specific stimulus dimension, two events occur. First, attributes of the attended dimensions become more dissimilar from each other. Second, attributes of the unattended dimensions become more similar to each other. Based on several categorization studies, Nosofsky (1986, 1987) has shown that a selective attention strategy to one dimension serves to maximize within-category similarity among exemplars sharing the same dimensions and minimize between-category similarity. Using this strategy, he was able to account for a number of seemingly diverse findings in the categorization-classification literature.

One consequence of this view of selective attention for speech perception is that it provides a way to account for the effects of linguistic experience quite easily in terms of modifications in the relative salience of different phonetically relevant dimensions depending on the specific language-learning environment. For example, in one study, Terbeek (1977) used a scaling technique to measure the magnitudes of differences between pairs of vowels presented to native speakers of five different languages. He found that prior language experience affects vowel perception by modifying the perceived psychological distances. The perceptual distance between a pair of physically similar vowels was judged to be much larger if members of the pair contrasted phonologically in the subject's native language than if the pair was not phonologically distinctive in the language. Thus, the effects of linguistic experience apparently modify the underlying psychological scales by altering the similarity relations for different perceptual dimensions. Jusczyk (this volume) has also made the same suggestions based on infant data.

According to this approach, linguistic experience affects perception by modifying attentional processes which, in turn, affect the underlying perceived psychological dimensions. When viewed in this context, the apparent developmental loss brought about by acquiring a language is not a "true" sensory-based loss but rather a change in selective attention. In principle, it should be possible to demonstrate that all non-native speech contrasts can be discriminated reliably by adults in a short period of time using relatively simple laboratory training techniques. Because the underlying sensory abilities are still intact, discrimination training only serves to modify attentive processes which are assumed to be flexible and susceptible to realignment (see Aslin & Pisoni, 1980).

If the changes in perceptual reorganization and the associated development loss in speech perception are primarily attentional in nature, then what are the consequences of this approach for discrimination and categorization of non-native speech contrasts? One consequence of this view is a systematic "warping" or restructuring of the psychological space, favoring important distinctive contrasts that are present in a particular language and the attenuation of cues for non-native distinctions. However, in addition to modifying the perceived psychological spacing among dimensions, there are also changes in the memory representations for the psychologically more salient dimensions. Changes in memory could account for the large and reliable differences observed in perception for within- versus between-category comparisons in speech discrimination tasks (Pisoni, 1973, 1975). Similarly, changes in memory representations could also account for why discrimination training tasks, such as the AX test which emphasizes very fine within-category acoustic differences, fail to produce reliable differences in subsequent categorization and identification tasks with real words (Strange & Dittman, 1984). Listeners have the underlying sensory mechanisms to make very fine phonetic discriminations but they cannot develop stable representations in long-term memory that can be used later on in other tasks that require more abstract memory codes (Pisoni, 1973).

One additional point about Nosofsky's work on selective attention is relevant here in terms of tests involving transfer and generalization of knowledge gained in training. Following earlier work of Tversky (1977), Nosofsky (1986, 1987) has shown that similarity relations for complex multidimensional stimuli are not invariant over tasks or contexts and that attentional processes may operate differently under different experimental procedures. According to this view, similarity is highly context-dependent and subjects may attend selectively to different dimensions when the cues are in different contexts.

This view of similarity should come as no surprise to anyone working in the field of speech perception. One of the most distinctive and pervasive characteristics of speech is its highly context-dependent nature. Given this view of attention and similarity, Strange and Dittman's (1984) failure to find any transfer of training from pretest to posttest performance is easy to understand. Indeed, the contexts used in training (i.e., synthetic "rock–lock") were so different from those used in testing (i.e., minimal pairs of natural speech) that only a context-independent learning strategy using highly abstract perceptual units like phonemes would have produced any positive transfer effects. The data from a large number of theoretical and experimental studies over the last 40 years show that while phonemes may be extremely useful for linguistic descriptions, their status in the real-time processing of spoken language is still problematic (see Pisoni & Luce, 1987).

Training and Perceptual Learning

While this chapter has been nominally concerned with training and perceptual learning effects in speech perception, the major focus has been on a number of general issues related to perceptual reorganization and developmental change and the mechanisms that underlie these processes. One of our major findings was that short-term laboratory training procedures do apparently produce changes in speech perception and these changes generalize to items that subjects were not originally trained on. Our findings are also consistent with those reported recently by Jamieson and Morosan (1986), who were also successful in modifying the speech perception abilities of naive adults with laboratory training techniques. The outcome of their training study on the perception of English fricatives by Canadian francophones, like the present results on /r/ and /l/, demonstrated several important methodological principles that should be followed to insure successful transfer of training. According to Jamieson and Morosan (1986), previous failures to train non-native speech contrasts reflect deficiencies in the specific training methods. They propose three training principles. The first principle deals with acoustic context. They argue that training should ensure that the relevant speech cues are presented in an acoustic context that is appropriate for normal speech rather than in isolation. The acoustic cues for segmental distinction in speech are highly context dependent. Removing a cue from context alters the perception of that cue when it is put back into the context again.

The second principle involves identification training. According to Jamieson and Morosan (1986), the training task should require subjects to make identification responses rather than discrimination responses because the desired outcome of training is to selectively modify the listener's ability to classify speech sounds into new perceptual categories. Training subjects in discrimination tasks will have the undesirable effect of enhancing perception of fine within-category acoustic differences rather than encouraging categorization responses.

Finally, training should begin by initially focusing attention on the criterial acoustic cues and then introducing a range of acoustic variability into the stimulus materials. If the goal of training is to improve perception of natural speech, listeners must eventually learn to ignore within-category acoustic variability while attending to relevant between-category variability.

Jamieson & Morosan (1986) argue that it will be difficult to train listeners to perceive many new speech contrasts without including some kind of stimulus variability in training so the listener will be able to eventually deal with the inherent variability encountered in naturally produced speech.

Segmentation, Perceptual Units and Language Development

The results of our /r/-/l/ training study also raise several theoretical issues about what adult subjects are actually learning in experiments like these. An analysis of both the training data and the pretest–posttest data for /r/ and /l/ demonstrate that subjects are learning about these phonetic contrasts in a highly context-dependent manner. We found very little evidence that subjects were "extracting" out or using context-independent perceptual attributes or units like traditional phonemes. If anything, the data demonstrate that subjects were encoding very specific contextual information about the acoustic cues for /r/ and /l/ in the different phonetic environments that they were exposed to during training. While these data are from mature adult listeners who have already acquired their native language, the present results nevertheless raise several important questions about what infants and young children are encoding from the speech in their environment. The available evidence to date suggests that infants and young children are not breaking down the speech signal into phoneme-like segments at all but rather are responding to much more global characteristics of the similarity structure of speech that they hear in their environment (Ferguson, 1986; Studdert-Kennedy, 1987).

Since the initial report by Eimas et al., (1971), there has been a great deal of speculation about precisely what the infant speech perception findings mean and how they may be related to later stages of language development. Some investigators such as Werker (Werker & Lalonde, 1988) and Best et al., (1988) believe that the discrimination data demonstrate that infants are perceiving speech signals in a "phonetically relevant" manner and that the underlying perceptual abilities will someday be useful for the eventual task of language learning. An example of this line of reasoning is given below:

> This mapping between biologically given sensitivities and phonetic categories allows the young infant to segment the incoming speech stream into discrete perceptual entities and enables the infant to divide the ongoing and overlapping stream of speech into the units that will be required in the important task of beginning to learn a language. (Werker & Lalonde, 1988, p. 681)

We believe there are some problems with these claims. Inherent in this line of speculation is the assumption that young infants somehow acquire the sound structure and lexicon of a language by a bottom-up strategy involving the segmentation of the acoustic signal into perceptual entities that correspond to adult-like units such as phones, allophones, phonemes and words. While such a sensory-based approach is appealing, it also lacks empirical support.

These speculations are also closely tied to a set of theoretical assumptions derived from an adult model of language development in which segments and features play an important theoretical role in studies of infant speech sound perception and early lexical development. However, the data from studies of child phonology suggest a very different picture of development. According to Studdert-Kennedy (1987), the child first uses relatively large undifferentiated units, such as prosodic contours, to express meanings. When the child begins to use sound patterns in a contrastive way, the units of production are more likely to approximate word-like patterns than smaller units such as segments. These early words appear to function as wholistic units that have no coherent internal structure. Many of the phonetic forms used in these first words are also highly context specific and suggest that the child has little, if any, awareness or active control over the arrangement and sequencing of the component sounds in these patterns.

Given what we currently know about the child's early attempts at speech production, it seems very unlikely that children are actively segmenting the incoming speech signal into relatively small perceptual units for subsequent identification and categorization into phonemes and then words. Instead, the perceptual process may be just the opposite as the child moves progressively from relatively large undifferentiated units, to smaller context-dependent units as his/her lexicon begins to increase in size (see Moskowitz, 1973; Studdert-Kennedy, 1986; Jusczyk, 1986). Only when the size of the child's lexicon becomes too large for efficient retrieval, will a segmentally-based production strategy begin to emerge. At this point, the child is able to control and coordinate articulatory gestures in speech production so that the sound patterns can function contrastively to express different meanings. Until the lexicon becomes organized, words and larger sound patterns are probably the units common to both perception and production.

Thus, viewed in this manner, the extensive literature on infant speech sound perception over the last 18 years merely provides information about the sensory capabilities of infants and not much more than that. As Studdert-Kennedy (1989) has remarked, the infant behaves as a "psychophysical chinchilla". The discrimination findings reported in the literature must therefore be interpreted cautiously with regard to any strong claims about their direct relevance to the development of the child's lexicon or the emerging functional use of spoken language as the child matures. It is clear from these studies that the young child can discriminate phonetically relevant acoustic contrasts, contrasts that will later become important in the language the child eventually develops. However, it is not at all clear precisely how these phonetic proclivities interface with the child's developing lexicon or his/her command of phonology during the first two or three years of life.

Conclusions

The findings reported in this chapter permit us to make several general conclusions about the effects of laboratory training on speech perception. First, based on the two studies reviewed here, there can be little question that laboratory training procedures can be used to selectively modify the perception of non-native phonetic contrasts. The limitations and apparent perceptual difficulties observed with non-native contrasts in the past appear to be due to selective attention and memory processes rather than any basic limitation on the sensory capabilities underlying these particular phonetic contrasts.

Second, the perception of /r/ and /l/ appears to depend quite extensively on the specific phonetic environment in which the contrast appears. In both training and testing, subjects displayed strong evidence that their encoding and subsequent internal representations of these sounds were highly context dependent. We found no evidence to suggest that subjects were attempting to encode these contrasts in terms of some abstract context-independent perceptual units like phonemes or phonetic segments.

Third, the learning and generalization effects that we observed apparently were facilitated by the high degree of stimulus variability used during the training phase. By exposing subjects to variability from different phonetic contexts and from different talkers, subjects apparently developed "robust" phonetic categories which facilitated transfer of this knowledge to new environments and new talkers.

Fourth, the use of nonsense syllables and highly controlled synthetic speech stimuli in past training studies may have placed a number of constraints on subjects' learning strategies. The success of the present /r/ and /l/ study can be attributed, in part, to the use of phonetically redundant naturally produced English words in which /r/ and /l/ appeared in a wide variety of different phonetic environments. Using these procedures, subjects were able to acquire knowledge about the range of variability these contrasts might assume in natural speech tokens.

Finally, the results of our training studies are compatible with the view that developmental change and associated perceptual reorganization in speech perception is due primarily to selective attention rather than any permanent changes in the underlying sensory or perceptual mechanisms. According to this view, selective attention to linguistically relevant phonetic contrasts operates by "warping" the underlying psychological distances. For speech contrasts that are distinctive in the language-learning environment, the psychological dimensions are stretched so that important phonetic differences become more salient; for speech contrasts that are non-distinctive, the psychological dimensions are shrunk so that unattended differences become more similar to each other. This view of the role of selective attention in speech perception can accomodate a wide variety of developmental and cross-language findings in the literature and provide a psychological basis for the mechanisms underlying perceptual change.

References

Abramson, A. & Lisker, L. (1970). Discriminability along the voicing continuum: Cross language tests. *Proceedings of the 6th International Congress of Phonetics Sciences* (pp. 569-573). Prague: Academia.

Aslin, R. N. (1981). Experiential influences and sensitive periods in perceptual development: A unified model. In R. N. Aslin, J. R. Alberts, & M. R. Peterson (Eds.), *Development of Perception: Psychobiological Perspectives, (Vol. II): The Visual System* (pp. 45-93). New York: Academic Press.

Aslin, R., N. (1985). Effects of experience on sensory and perceptual development: Implications for infant cognition. In J. Mehler & R. Fox (Eds.), *Neonate Cognition: Beyond the Blooming, Buzzing Confusion* (pp. 157-183). Hillsdale, NJ: Erlbaum.

Aslin, R. N. (1987). Visual and auditory development. In J. D. Osofsky (Ed.), *Handbook of Infancy, 2nd Edition*. New York: Wiley.

Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child Phonology: Perception and Production* (pp. 67-96). New York: Academic Press.

Aslin, R. N., & Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, **52**, 1135-1145.

Aslin, R. N., & Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in early infancy. In M. Haith & J.Campos (Eds.), *Handbook of Child Psychology, Vol. 2, Infancy and Developmental Psychobiology* (pp. 573-687). New York: Wiley.

Best, C. T., MacRoberts, G. W., & Sithole, N. M. (1988). Examination of the perceptual re-organization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 245-260.

Burnham, D. K. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. *Applied Psycholinguistics*, 7, 207-240.

Carney, A., Widin, G., & Viemeister, N. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*,**62**, 961-970.

Dissosway-Huff, P., Port, R., & Pisoni, D. B. (1982). Context effects in the perception of /r/ and /l/ by Japanese. *Research on Speech Perception Progress Report No. 8.* Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*,**46**, 372-383.

Eilers, R. E., Gavin, W. J., & Wilson, W. R. (1979). Linguistic experience and phoneme perception in infancy: A cross-linguistic study. *Child Development*, **50**, 14-18.

Eimas, P.D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303-306.

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics*, **18**, 341-347.

Eimas, P. D. (1978). Developmental aspects of speech perception. In R. Held, H. Leibowitz, & H. L. Teuber (Eds.), *Handbook of Sensory Physiology: VIII. Perception* (pp. 357-374). New York: Springer-Verlag.

Ferguson, C. A. (1986). Discovering sound units and constructing sound systems: It's child's play. In J. Perkell & D. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 36-51). Hillsdale, NJ: Erlbaum.

Flege, J. E. (1987). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human Communication and its Disorders, Vol. I* (pp. 224-401). Norwood, NJ: Ablex.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers in Linguistics and the Philosophy of Language*, **6**, 59-72.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, **9**, 317-323.

Gottlieb, G. (1981). The roles of early experience in species-specific perceptual development. In R. N. Aslin, J. R. Alberts, & M. R. Peterson (Eds.), *Development of Perception: Psychobiological Perspectives, Vol. 1, Audition, Somatic Perception, and the Chemical Senses* (pp. 5-44). New York: Academic Press.

Jamieson, D. & Morosan (1986) Training non-native speech contrasts in adults: Acquisition of the English /o/-/0/ contrast by francophones. *Perception & Psychophysics*, **40**, 205-215.

Jenkins, J.J. (1989) The more things change, the more they stay the same: Comments from an historical perspective. In R. Kanfer, P.L. Ackerman and R. Cudeck (Eds.), *Abilities, Motivation, and Methodology* (pp. 475-491). Hillsdale, NJ: Erlbaum.

Jusczyk, P. (1985). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the Blooming Buzzing Confusion* (pp. 199-229). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W. (1986) Toward a model of the development of speech perception. In J. Perkell & D. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 1-19). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W. (1990; this volume). Infant speech perception and the development of the mental lexicon. To appear in H. C. Nusbaum & J. C. Goodman (Eds.), *The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception* (pp. 00-00). Cambridge, MA: MIT Press.

Jusczyk, P., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L., & Mehler, J. (1989). *The role of attention in speech perception by young infants.* Manuscript submitted for publication.

Kuhl, P. K. (1987). Perception of speech in early infancy. In P. Salapatek & L. Cohen (Eds.) *Handbook of Infant Perception, Vol. 2* (pp. 275-381). New York: Academic Press.

Kuhl. P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, **190**, 69-72.

Lane, H. L. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, **72**, 275-309.

Lasky, R. E., Syrdal-Lasky, A., & Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, **20**, 215-225.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Liberman, A. M., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.

Lisker, L., & Abramson, A. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Lisker, L., & Abramson, A. (1967). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563-567). Prague: Academia.

Lisker, L. (1970). On learning a new contrast. *Status Report on Speech Research (SR-24)*. New Haven, CT: Haskins Laboratories

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1988). Training Japanese listeners to identify /r/ and /l/: A first report. *Research on Speech Perception Progress Report No. 14*. Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

McClasky, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics*, **34**, 323-330.

MacKain, K., Best, C., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, **2**, 369-390.

Mann, V. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of “l' and "r“. *Cognition*, **24**, 169-196.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **15**, 676-684.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception & Psychophysics*, **18**, 331-340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics*, **9**, 283-303.

Moskowitz, A. I. (1973). The acquisition of phonology and syntax. In G. Hintikka, J. Moravesik, and P. Suppes (Eds.), *Approaches to Natural Language* (pp. 48-84). Dordrecht: Reidel.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Nosofsky, R. M.(1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **14**, 700-708.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In H. C. Nusbaum & E. C. Scwab (Eds.), *Pattern Recognition by Humans and Machines, Vol. 1: Speech Perception* (pp. 113-157). New York: Academic Press.

Pisoni, D. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.

Pisoni, D. B. (1975). Auditory short-term memory vowel perception. *Memory &Cognition*, **3**, 7-18.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Ed.), *Handbook of Learning and Cognitive Processes, Vol. 6* (pp. 167-233). Hillsdale, NJ: Erlbaum.

Pisoni, D., Aslin, R., Perey, A., & Hennessy, B. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 297-314.

Pisoni, D. & Lazarus, J. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

Pisoni, D. & Luce, P. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, **25**, 21-52.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, **15**, 285-290.

Sheldon, A. & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, **3**, 243-261.

Shiffrin, R. M. (1976). Capacity limitations in information processing, attention, and memory. In W. K. Estes (Ed.), *Handbook of Learning and Cognitive Processes, Vol. 4* (pp. 177-236). Hillsdale, NJ: Erlbaum.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human Communication: A Unified View* (pp. 51-66). New York: McGraw-Hill.

Stevens, K. N. (1980) Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, **68**, 836-842.

Strange, W. (1972). *The Effects of Training on the Perception of synthetic Speech Sounds: Voice Onset Time.* Unpublished doctoral dissertation, University of Minnesota.

Strange, W. & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, **36**, 131-145.

Strange, W. & Jenkins, J. (1978). Role of linguistic experience in the perception of speech. In R. D. Walk & H. L. Pick (Eds.), *Perception and Experience* (pp. 125-169). New York: Plenum Press.

Streeter, L. A. (1976a). Kikuyu labial and apical stop discrimination. *Journal of Phonetics*, **4**, 43-49.

Streeter, L. A. (1976b). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature*, **259**, 39-41.

Studdert-Kennedy, M. (1986). Sources of variability in early speech development. In J. Perkell & D. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 58-76). Hillsdale, NJ: Erlbaum.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scherer (Eds.) *Language Perception and Production* (pp. 67-84). London: Academic Press.

Tees, R., & Werker, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology*, **34**, 579-590.

Terbeek, D. A. (1977). A cross-language multi-dimensional scaling study of vowel perception. *Working Papers in Phonetics* (University of California at Los Angeles), **37**.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Walley, A., Pisoni, D. B., & Aslin, R. N. (1981). The role of early experience in the development of speech perception. In R. N. Aslin, J. R. Alberts, & M. R. Peterson (Eds.), *Development of perception: Psychobiological perspectives, Vol. 1, Audition, Somatic Perception, and the Chemical Senses* (pp. 219-255). New York: Academic Press.

Werker, J. F. (1989, January-February). Becoming a native listener. *American Scientist*, **77**, 54-59.

Werker, J. F., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, **24**, 672-683.

Werker, J. F. & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866-1878.

Modes of Processing Speech and Nonspeech Signals[1]

David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] **This is a draft of a paper presented at the Conference on Modularity and the Motor Theory of Speech Perception, June 5-8, 1988, New Haven, Connecticut. Preparation of this paper was supported by NINCDS Research Grant NS-12179 to Indiana University.**

Modes of Processing Speech and Nonspeech Signals

I am delighted to be here today to honor Al Liberman and talk about speech perception, particularly some findings that I know Al would be happy to hear about given the theme of the conference on Motor Theory and Modularity. I feel a little awkward in talking to this audience about some of these findings because I find myself in the rather uncomfortable position of disagreeing with the major conclusions of Al Bregman's talk, on the one hand, and defending Al Liberman's views about speech on the other hand. This may seem a little curious to most of you in the audience since I have always thought of myself as the "quiet" little gadfly off working in my Lab back in Indiana in the heartland of America far away from Haskins Laboratories and all the day to day excitement that goes on there.

After reading Al Bregman's paper and thinking about his major claims, I began looking through my slide collection to find something to say about his arguments. As I thought more and more about what he said about speech perception and auditory perception, particularly "auditory scene analysis," I began to realize that maybe Al Liberman's arguments for the "preemptiveness" of speech might *not* be as "bizarre" as I once thought they were when I first read the Liberman and Mattingly (1985) *Cognition* paper a few years ago. When you think about things one way for a long time, it is often very difficult, if not impossible, to see them any other way! Perhaps it takes a paper like Al Bregman's to get one to think a little like Al Liberman and to appreciate some of the very deep and unusual claims that Al has made over the years about speech.

In order to see how my thinking went over the last few weeks, I will begin by going back about 25 years to the days of the Pattern Playback in the 1950s and 1960s and briefly review some of the now "classic" experiments in speech perception that form the knowledge base and early history of our field (see Figure 1).

Insert Figure 1 about here

For a considerable number of years, Al Liberman and his collaborators at Haskins Laboratories have been interested in the differences in perception between speech and nonspeech signals. That such differences might exist was suggested by the report of the very first findings on categorical perception of stop consonants back in 1957 (Liberman, et al., 1957). As is now well-known to everyone, even introductory psychology students, the discrimination of most nonspeech continua is continuous and monotonic with changes in the physical scale. Observers are able to discriminate many more differences than they can reliably identify on an absolute basis. However, in the case of categorical perception, listeners can discriminate between two stimuli no better than they can identify them as different on an absolute basis, suggesting that discrimination of speech was in some way limited by absolute identification.

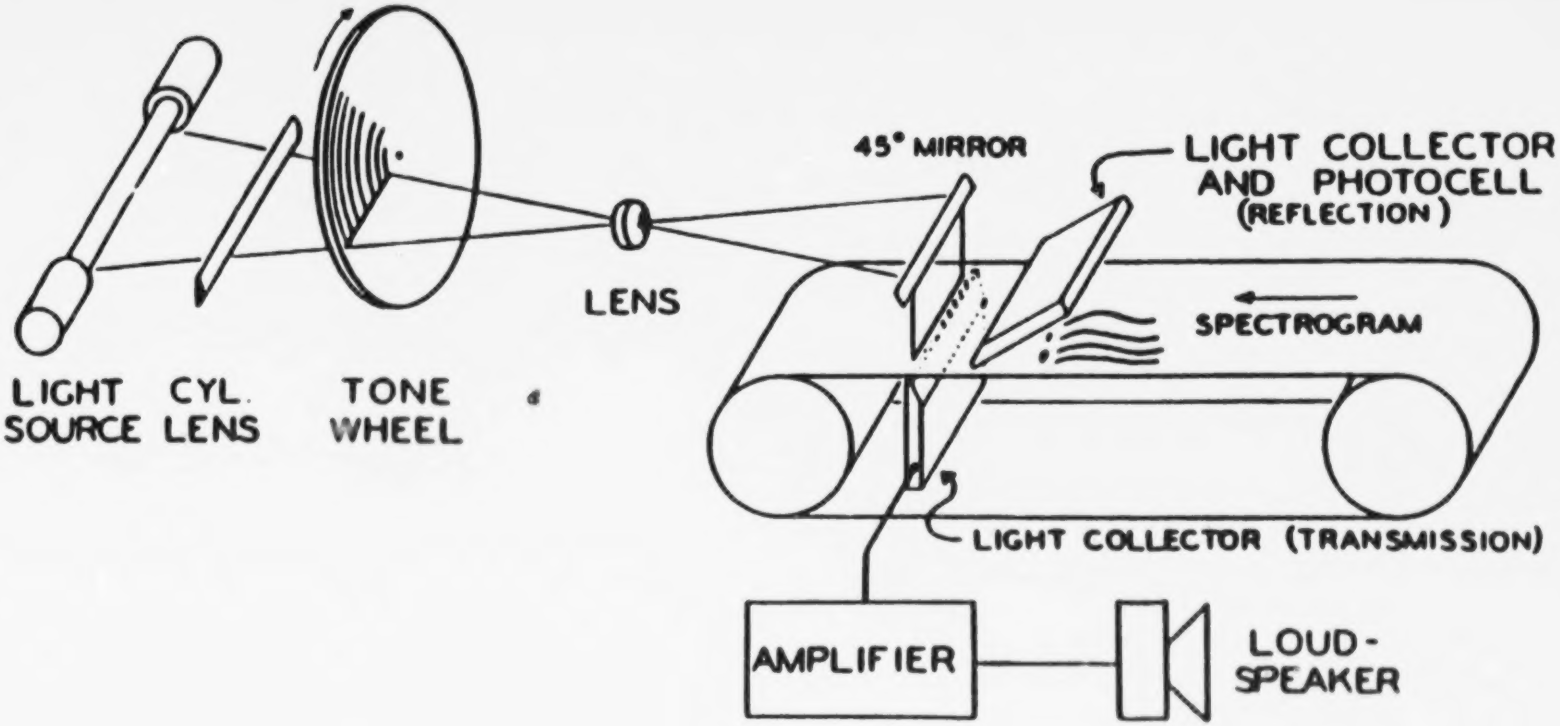

LIGHT SOURCE
CYL. LENS
TONE WHEEL
LENS
45° MIRROR
LIGHT COLLECTOR AND PHOTOCELL (REFLECTION)
SPECTROGRAM
LIGHT COLLECTOR (TRANSMISSION)
AMPLIFIER
LOUD-SPEAKER

Figure 1. A schematic representation of the Haskins Laboratories Pattern Playback speech synthesizer. Simplified spectrographic patterns of speech were painted by hand on a moving acetate belt to represent the important time-varying formant structure of an utterance. The device would convert these visual patterns into highly intelligible speech that could be used in perceptual experiments with human listeners. Photographs of natural spectrograms could also be used as input to the device in order to reproduce utterances in synthetic form (From Cooper, 1950).

Since, at least at that time in the late 1950s, categorical perception seemed to be restricted to speech stimuli, particularly the perception of stop consonants, it became of some interest to determine the underlying basis for the non-monotonic discrimination functions found for speech as shown by the excellent discrimination observed at category boundaries and the relatively poor discrimination observed within categories.

It was with this general goal in mind that the first "nonspeech control" experiment was carried out by Al and his colleagues (Liberman et al., 1961). In this study, Liberman et al. (1961) wanted to determine whether the peaks in the discrimination function observed for stop consonants were a result of learning or whether they were given innately. If the peaks in discrimination could be attributed to learning, then an additional concern was to determine precisely what kind of learning was involved in the process. In order to answer both of these questions, Liberman et al. (1961) created a set of nonspeech stimuli by inverting the synthetic spectrographic patterns appropriate for the /do/-/to/ continuum before converting them to sound on the pattern playback. The aim of this manipulation was, at least in principle, to generate a set of nonspeech control stimuli that had all the properties of the speech stimuli but actually did not sound like speech. Examples of these stimuli are shown in Figure 2.

Insert Figure 2 about here

The strategy of the Liberman et al. (1961) study was then to compare the discrimination functions for the speech signals with those obtained for their nonspeech controls in order to ascertain whether the nonspeech signals would show comparable peaks and troughs in discrimination. At least two outcomes were possible from such an experiment. First, if discrimination peaks are present for the nonspeech stimuli and they occurred in roughly the same regions as those found for the speech condition, the findings could then be attributed to the specific acoustic properties of the signals themselves and not to any additional interpretative process whereby the signals were identified or encoded as speech. Liberman et al. (1961) suggested that this particular result would support an account of speech perception in terms of innate factors presumably involving some sort of psychophysical explanation.

The second possible outcome was quite different. Liberman et al. (1961) argued that if the discrimination peaks were absent from the nonspeech control stimuli, then a learning account would be appropriate. Moreover, depending upon the overall level of discrimination observed in the nonspeech condition, one of two possible learning explanations would be possible. Both of these learning explanations may be contrasted for the idealized cases as shown graphically in Figure 3.

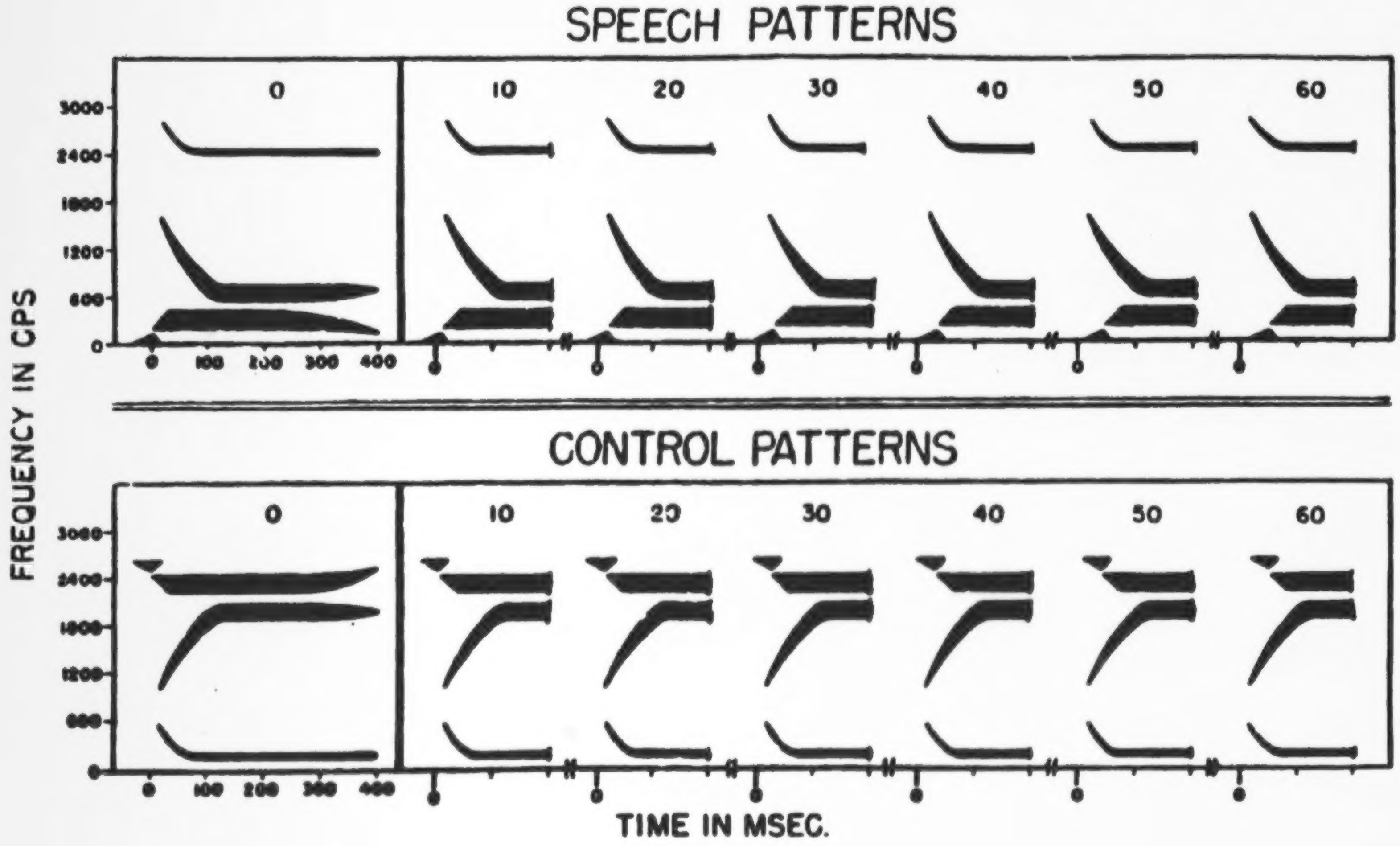


Figure 2. Top panel shows schematized spectrographic patterns appropriate for generating a continuum of synthetic speech stimuli ranging from /do/ to /to/. The bottom panel shows the set of control patterns that were created by inverting the speech patterns shown in the top panel. The control patterns contained all of the timing information present in the original speech patterns although they did not sound like speech (From Liberman, Harris, Kinney & Lane, 1961).

Insert Figure 3 about here

In both panels of this figure, a hypothetical ABX discrimination function for a selected speech continuum is shown by the solid lines and filled circles; the functions for the nonspeech control stimuli are shown as the dashed lines and open circles. Panel (a), on the left, illustrates the results predicted by a learning interpretation of the discrimination peak in terms of the concept of "acquired similarity." According to this view, the peak in the speech discrimination function arises from learning to *ignore* variations within phonological categories and learning to *attend* to variations across categories. Thus, discrimination of the relevant acoustic parameter underlying the speech contrast was originally presumed to be quite good, as shown by the high level of the nonspeech discrimination function. The effects of perceptual learning, therefore, serve to *attenuate* sensitivity selectively.

In contrast, Panel (b), on the right, illustrates the results predicted by an interpretation of the discrimination peak in terms of the concept of "acquired distinctiveness." According to this view, which was widely held by many psychologists back in the 1950s (see Gibson & Gibson, 1955), the peak in the discrimination function arises from learning to respond to stimuli that have somehow become more distinctive or salient to the listener through a process of *differentiation*. By this account, discrimination was originally assumed to be quite poor and the effects of perceptual learning were to make certain stimuli more *distinctive* by increasing the organism's sensitivity to them through exposure and feedback.

The results of the Liberman et al. (1961) /do/-/to/ control study did not reveal a peak in the ABX discrimination functions for the nonspeech stimuli, suggesting a learning explanation. In addition, the overall level of the discrimination function was quite low, very close to chance, suggesting that the peaks in the speech discrimination function were more likely to be the result of acquired distinctiveness than acquired similarity. At the time, Liberman et al. (1961) assumed that the distinctiveness for speech arose during the course of language learning through mediation of the production system in the process of learning the relevant articulatory gestures needed to produce the same distinction, a position characterizing what would eventually become one of the earliest statements of the "Motor Theory of Speech Perception." In the 1960s, these nonspeech results were often cited not only as evidence for the presence of important differences in perception between speech and nonspeech signals, but also as additional support for the view that speech perception might have very close ties with speech production (Liberman, Cooper, Harris & MacNeilage, 1963).

Numerous other speech-nonspeech comparisons have been carried out over the years at Haskins. All of these studies have revealed quite similar results, particularly with regard to comparisons involving the shape and relative level of the discrimination function (see Mattingly et al., 1971; Miyawaki et al., 1975). The nonspeech control signals have uniformly and consistently failed to show peaks in discrimination that were correlated with the peaks

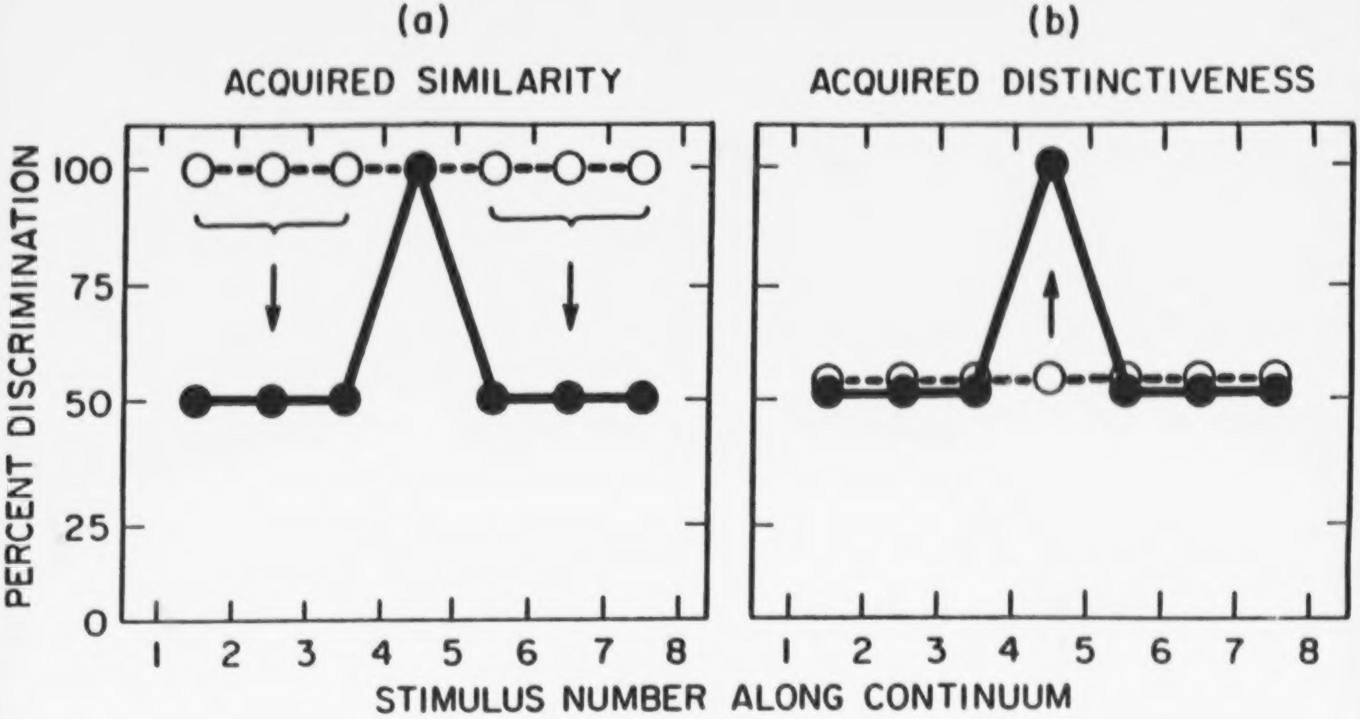


Figure 3. Idealized versions of ABX discrimination functions that would be expected from either the processes of acquired similarity (Panel a on the left) or acquired distinctiveness (Panel b on the right). These two functions attempt to illustrate how early linguistic experience might selectively modify speech discrimination capabilities.

in the speech identification functions. Moreover, these nonspeech stimuli were typically discriminated at levels approaching chance responding. In more recent years, Liberman and his colleagues have tended to avoid explanations of these differences in terms of the older notions of acquired similarity and acquired distinctiveness, preferring instead to characterize the differences in perception between speech and nonspeech as reflecting two basically different modes of perception, a speech mode and a nonspeech mode (Liberman, 1970). Until recently, the nonmonotonic discrimination functions observed for speech stimuli have typically been accounted for in terms of some additional, perhaps specialized, interpretative process involving phonetic categorization rather than a purely sensory-based process involving responses to only the psychophysical properties of the signals themselves (see Pisoni, 1977).

A number of criticisms can be leveled at these speech-nonspeech comparisons, particularly the results from the nonspeech control conditions. First, there is the question of whether the same psychophysical properties found in the original speech stimuli are indeed preserved in the nonspeech control stimuli. This criticism is appropriate for the original /do/–/to/ stimuli and the Mattingly et al. (1971) "chirp" and "bleat" controls where the acoustic cues were removed from speech context and presented in isolation. This manipulation, while nominally preserving the speech cue, results in a marked change in the spectral context which no doubt affects processing of the speech cue itself (see Stevens, 1980).

A second problem with these nonspeech control stimuli concerns the fact that subjects did not receive any experience or familiarization with these signals prior to the discrimination test. With complex multidimensional signals, it may be difficult for subjects to attend to the relevant attributes that distinguish different stimuli. Thus, a subject's performance may be no better than chance if he is not attending selectively to the specific criterial attributes that distinguish these stimuli. Since all of the early nonspeech experiments were also run without feedback, listener may have focused their attention on one aspect or set of attributes on a given trial and an entirely different aspect of the stimulus on the next trial. As a result, listeners may have responded to the same stimulus quite differently at different times during the course of the experiment thus revealing a level of performance no better than chance, which is precisely what Liberman et al. found in their early studies.

Finally, subjects in almost all of these nonspeech experiments never overtly labeled or identified these nonspeech stimuli into discrete perceptual categories before discrimination was measured, as is commonly done in the speech experiments. The prior labeling experience may tend to emphasize some aspects of the stimulus pattern and attenuate others in selective ways not known to the investigator. Some of these criticisms were specifically taken into account in the more recent nonspeech experiments, which may have been responsible for their more successful outcome compared to the earlier studies (see Pisoni, Carrell & Gans, 1983).

Despite the methodological criticisms that can be leveled at these early nonspeech control experiments, a number of very general issues were identified at the time by Al Liberman and his colleagues, issues that still continue to occupy researchers today. Although discussion of concepts like "acquired distinctiveness" and "acquired similarity" in speech perception have faded away, there is currently a great deal of interest in the role of early linguistic experience in the development of speech perception in infants and young children and the effects of linguistic knowledge on speech perception in adults (see Aslin & Pisoni, 1980; Aslin et al., 1981). Many of the theoretical issues touched upon in these early Haskins papers are still topics of great interest, although they have been modified to accommodate recent developments in cognitive psychology, linguistics, and neurobiology.

A period of some ten years elapsed between the Liberman et al. (1961) /do/-/to/ study and the first report by Peter Eimas and his colleagues on the discriminative abilities of young infants (Eimas, Siqueland, Jusczyk, & Vigorito, 1971). This well-known study not only showed categorical-like discrimination performance in infants, but it also demonstrated that infants are able to make very fine discriminations of relevant speech cues at an early age. In the years following this pioneering study, a great deal of data have been obtained in infant studies that suggest that the form of learning in speech perception is probably more nearly one of "acquired similarity" rather than "acquired distinctiveness," at least with regard to the discrimination of speech sounds. Interest in the "loss" of discriminative abilities in language learning and the nature of the perceptual mechanisms underlying this loss is, of course, a topic of current interest and a great deal of research (Strange & Jenkins, 1978; Pisoni et al., 1982). Many people are now working on this problem with adults, infants, young children, and animals as well, following up on ideas and suggestions that Al Liberman made years ago.

Not only has there been recent interest and research on the nature and time-course of the loss of discriminative abilities in infants and young children, but research has also continued in several directions on the perception of nonspeech signals having properties that are similar to speech. While the specific issues have changed somewhat since 1961, many of the fundamental theoretical questions still remain the same today. It is hard to think back to 1961 and imagine if anyone thought about the impact and importance that the /do/-/to/ study would have for the future of research in speech perception, or for the interesting directions this work would take in the years to come. Indeed, it is rare in the field of experimental psychology for any set of issues to last for more than a few years. However, in the case of speech perception, and, more specifically, with regard to differences in perception between speech and nonspeech signals, the fundamental questions have apparently endured for more than 25 years and, what is more surprising, is that they are still prominent in current theoretical discussions today. Perhaps these issues have survived so long because they deal with very deep and fundamental problems of perception and knowledge and their relation to language. Perhaps they have endured simply because they have not as yet received any satisfactory theoretical account. Who really knows? Regardless of the final explanation, research continues on problems that were first identified by Al Liberman back in 1961 using similar methodologies and experimental designs. Some of these recent findings have revealed important new properties about speech which lead some researchers to suppose that biologically specialized mechanisms are needed for perceptual analysis (see Liberman and Mattingly, 1985).

In addition to several prominent differences in the acoustic characteristics of speech and nonspeech sounds which set speech signals apart from other auditory signals in a listener's auditory environment, there are also a number of distinctive differences in the way in which speech and nonspeech sounds are encoded, recognized, and identified. Research by Al and his colleagues at Haskins has demonstrated that when human listeners are presented with speech signals, they typically respond to them as linguistic entities rather than as isolated auditory events in their environment. The set of labels used in responding to speech is not arbitrary – the labels are intimately associated with the function of speech as the signalling system used in spoken language. Speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background and experience. Moreover, a listener's performance in identifying and discriminating a particular acoustic attribute is often a consequence of the functional role this property plays in the listener's own linguistic system. In some of my own studies, we have shown that it is possible to get human listeners to respond to the auditory properties of speech signals and to "hear out" certain components with extensive training and the use of very sensitive psychophysical procedures (Pisoni & Lazarus, 1974; Pisoni & Tash, 1974). But one of the fundamental differences in perception between speech and nonspeech sounds lies in the linguistic significance of the stimulus patterns to the listener and the context into which these patterns are subsequently integrated.

One very clear example of differences in mode of processing comes from a study carried out at Indiana by Mary Ellen Grunke and I several years ago on the perception of complex nonspeech auditory patterns that have properties that are similar to speech (see Grunke & Pisoni, 1982). Subjects were required to identify auditory patterns with either acoustic or phonetic labels. No feedback was provided in this experiment, since we wanted to measure subjects' ability to categorize these auditory patterns solely on the basis of the acoustic or phonetic attributes implicit in these signals. Thus, we were not interested in the subjects' ability to learn an arbitrary sound-to-label association in the context of a particular test situation, as we have done in some of our previous experiments. The stimulus patterns were the single-, double-, and triple-tone signals shown in Figure 4. Two conditions were examined. In the "phonetic" condition, subjects were told that the stimuli were distorted tokens of natural speech. The response labels provided to subjects were the syllables "ba," "da," "ab," and "ad," which were placed under four separate buttons on a response panel. In the "acoustic" condition, the subjects were told that the stimuli were frequency-modulated tones generated by a computer and that they consisted of a short interval with constant pitch, preceded or followed by a very rapid rise or fall in pitch. The response labels were schematic line drawings of the time course of the frequency change of each stimulus, /—, —/, \—, —\.

Insert Figures 4 and 5 about here

Responses were scored as correct or incorrect depending on whether the indicated label was the most appropriate cue for the presented stimulus. Percent correct performance for both labeling conditions across the three stimulus sets is displayed in Figure 5. For the single- and double-tone stimuli, the subjects were able to use acoustic labels more accurately than phonetic labels (single tones: 49.8% vs. 36.6% correct for acoustic and phonetic labels, respectively; double tones: 61.0% vs. 42.2%). However, with the triple tones, which contained energy in the first formant region, listeners assigned phonetic labels much more accurately than acoustic labels (62.7% correct for phonetic labels compared with 42.6% for acoustic labels). Subjects in the phonetic labeling condition were apparently able to hear these triple-tone patterns as speech, whereas subjects in the acoustic labeling condition had much more difficulty in focusing their attention on the individual components of the patterns. The decrement in performance for the acoustic labeling group can be accounted for by the presence of conflicting information in the F1 transition region for half the stimuli. In the triple-tone patterns, the F1 always rises in initial position and falls in final position. For stimuli containing rising transition in initial position (i.e., /ba/) or falling transitions in final position (i.e., /ab/), the information in F1 is *correlated* and *redundant* with the direction of the movements of F2 and F3, thus facilitating performance. However, for stimuli containing falling transitions in initial position (i.e., /da/) and rising transitions in final position (i.e., /ad/), the F1 component conflicts with the direction of the transitions in the remainder of the pattern. Thus, listeners could attend to either the phonetic or acoustic properties of these signals with better-than-chance accuracy, but their overall level of performance varied with the complexity of the signal and the specific stimulus properties that were attended to under the two labeling conditions.

Examination of the labeling performance for the four separate stimuli shown in Figure 6 indicated that, for acoustic labels, response accuracy was much greater for transition-final signals ("ab," 69.44%; "ad," 62.83%) than for transition-initial signals ("ba," 38.94%; "da," 33.39%). Interestingly, however, when listeners assigned phonetic labels to these same signals, the differences between transition-initial and transition-final signals were reduced substantially and did not differ significantly from each other ("ba," 45.17%; "da," 43.00%; "ab," 51.28%; "ad," 49.27%). The data shown in Figure 6 have been pooled across single-, double-, and triple-tones, since the observed overall pattern of responses was essentially the same for each stimulus set. These results demonstrate a very marked dissociation in perception between auditory and phonetic categorization of the *same* acoustic signals.

Insert Figure 6 about here

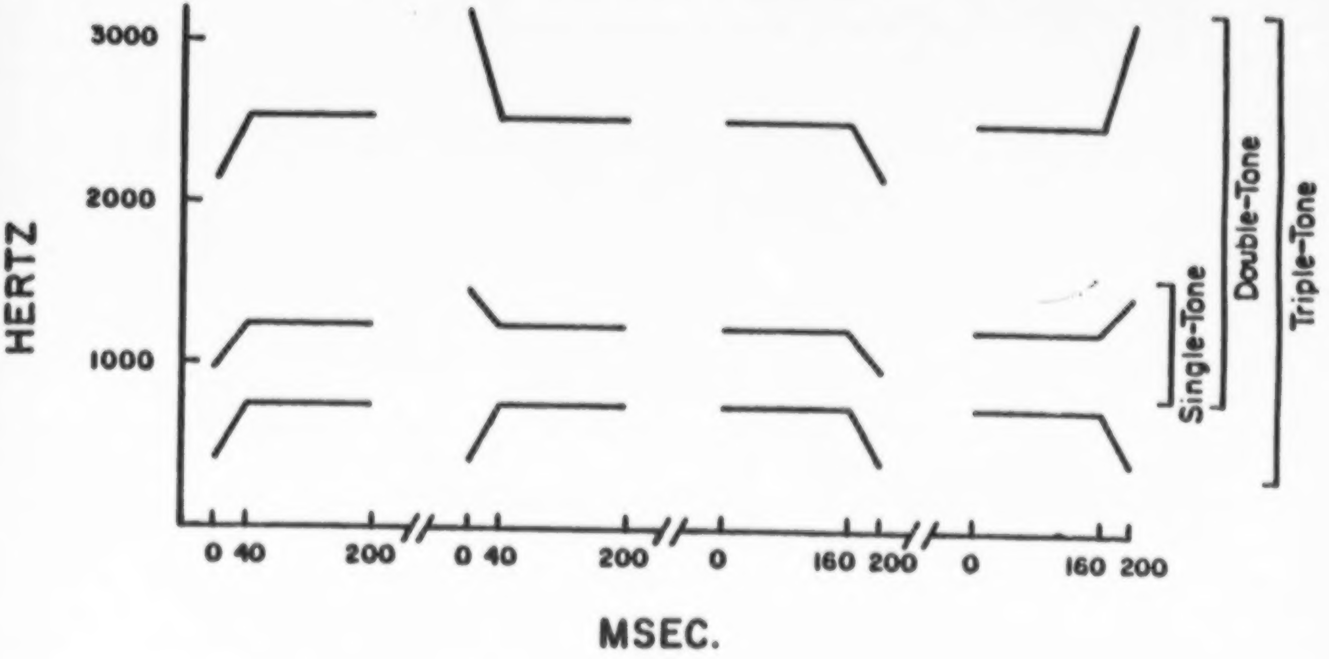

3000
2000
1000
HERTZ
0 40 200
0 40 200
0 160 200
0 160 200
MSEC.
Single-Tone
Double-Tone
Triple-Tone

Figure 4. Schematic spectrographic patterns of the four nonspeech stimuli used in the Grunke and Pisoni (1982) experiments.

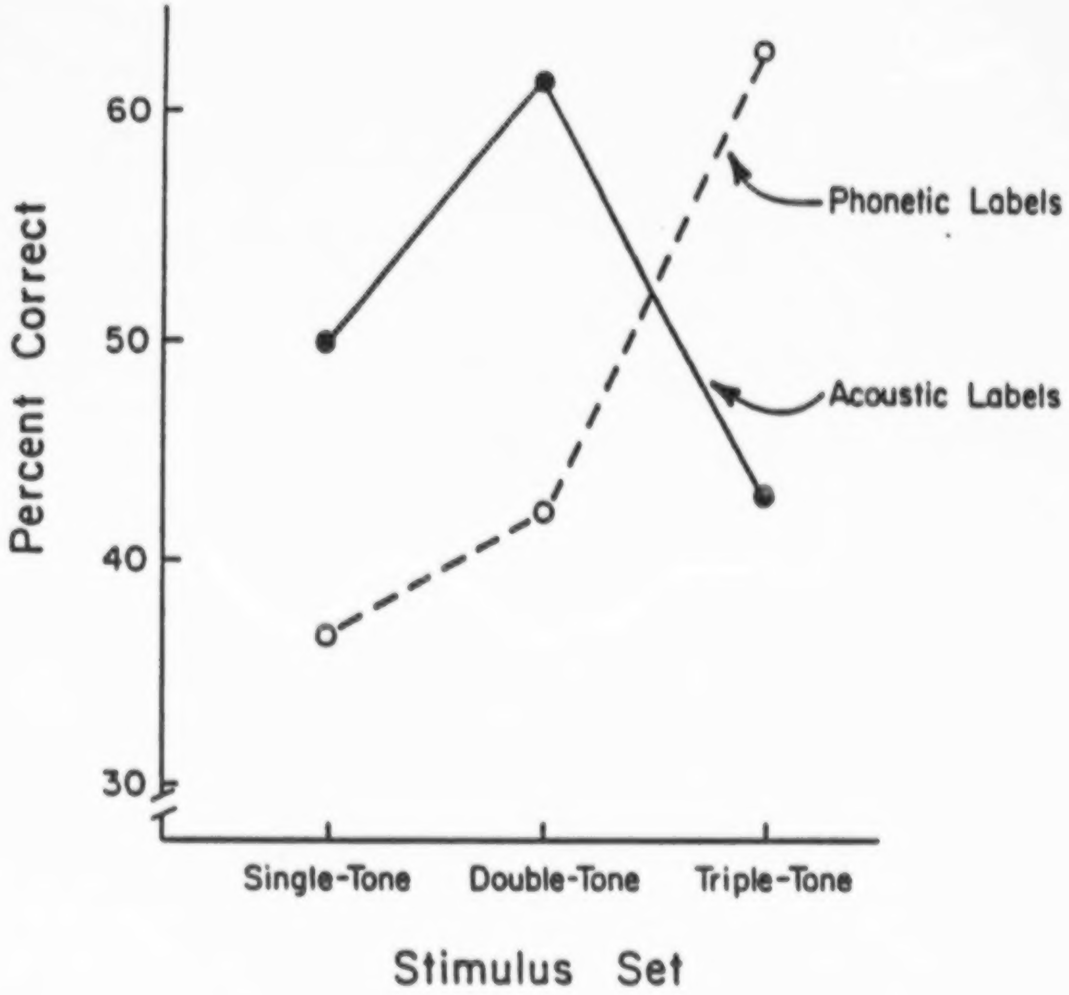


Figure 5. Percent correct identification in the labeling task for acoustic and phonetic labels. The data are shown separately for single-, double-, and triple-tone stimuli for each labeling condition.

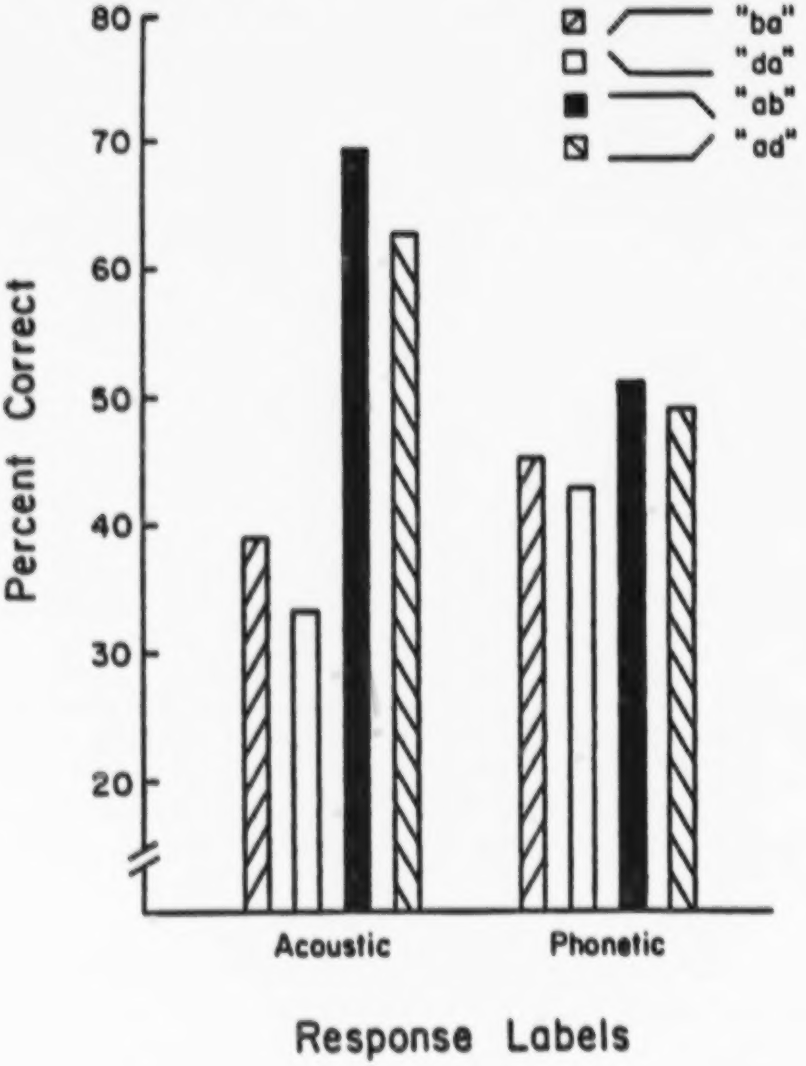


Figure 6. Percent correct identification for each of the individual stimuli used in the labeling task. Data are shown separately for acoustic and phonetic labeling conditions.

Two other related findings obtained in Jim Sawusch's Lab at Buffalo have also demonstrated marked differences in perception between speech and nonspeech signals as a function of mode of processing. In one study, Eileen Schwab (1981) found substantial backward masking effects and upward spread of masking for sine-wave stimuli heard as nonspeech tonal patterns. However, both masking effects were subsequently eliminated when the same identical sine-wave patterns were heard by listeners as speech. And, in a more recent study by Tomiak, Mullennix, & Sawusch (1987), using a Garner speeded classification task, subjects displayed evidence of processing separable dimensions with a set of noise-tone analogs of fricative-vowel syllables when they were told the patterns were nonspeech sequences. Irrelevant variation in the noise spectra *did not* affect reaction times for the classification of the tones and vice versa. However, when subjects were told the patterns were fricative-vowel speech stimuli, the component dimensions were processed in an integral manner such that irrelevant variation in the fricative increased the reaction times to classify the vowels and vice versa. These results were obtained with a set of noise-tone analogs that did not contain any consonant-vowel transitions from the noise into the steady-state segments of the patterns. Thus, for the noise-tone analogs heard as speech, knowledge of coarticulation between adjacent segments appears to have been used even when the acoustic cues for coarticulation were absent from the stimulus pattern. Tomiak et al. suggest that the use of knowledge about coarticulation in speech perception is "mandatory", in Fodor's sense, and is automatically invoked whenever an auditory pattern is heard as speech and processed in the speech mode (Fodor, 1983).

It is clear from these three sets of results and other recent studies carried out at Haskins using nonspeech signals which have properties similar to those found in speech that differences in "mode of processing" can control *perceptual selectivity* quite substantially and can subsequently influence the perception of individual components of the stimulus pattern as well as the entire pattern itself. This can occur in quite different ways with the same identical stimulus patterns, depending primarily on whether the subject's attention in the task is directed toward coding either the auditory properties of the signals or the phonetic content of the overall patterns. In the former case, the process is more analytic, involving the processing or "hearing out" of the individual components of the stimulus, whereas in the latter case, the process is more nearly "holistic", insofar as the individual components may be combined to form well-defined and highly familiar perceptual (phonetic) categories. With regard to the perception of speech, these results imply that listeners probably do not isolate and then subsequently process only the distinctive speech cues in the stimulus. Rather, it seems very likely that listeners respond to these so-called "speech cues" as simply part of the configuration of a spectrally complex dynamic time-varying auditory pattern. In the case of speech, the patterns have certain well-defined distinctive properties and display spectral-temporal relations that elicit a qualitatively different mode of processing, a speech mode, that appears to be quite different from the way other nonspeech auditory signals are responded to under similar conditions. Such findings are probably not too surprising to anyone sitting in the audience today. But in thinking about these results and what they imply for theories of speech perception, we should not forget that it was Al Liberman who first raised these same questions more than 25 years ago in the now classic /do/-/to/ experiment and it was Al Liberman who has relentlessly continued to pursue these and many other difficult and challenging problems in speech perception ever since. Perhaps some of us will get around to working on answers to the four questions that Al raised at the beginning of the conference yesterday. Let's just hope that it doesn't take us another twenty-five years to appreciate the many insights that Al has shown us about speech and language over the years. When this conference is over, maybe we can just go back to the Lab with Al again and get started working on some new experiments like many of us did years ago when we first learned about speech from Al himself.

References

Aslin, R.N. and Pisoni, D.B. (1980). Some developmental problems in speech perception. In G. Yeni-Komshian, J.F. Kavanagh, and C.A. Ferguson (Eds.) *Child phonology: Perception & production*. NY: Academic Press.

Aslin, R.N., Pisoni, D.B., Hennessy, B.L., and Perey, A.J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, **52**, 1135-1145.

Eimas, P.D., Siqueland, E.R., Jusczyk, P.W., Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303-306.

Fodor, J. (1983). *The modularity of mind*. Cambridge: MIT Press.

Gibson, J.J. and Gibson, E.J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, **62**, 32-41.

Grunke, M.E. and Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, **31**, 210-218.

Liberman, A.M. (1970). Some characteristics of perception in the speech mode. In D.A. Hamburg (Ed.), *Perception and its disorders, proceedings of A.R.N.M.D.* Baltimore: Williams and Wilkins. Pp. 238-254.

Liberman, A.M., Cooper, F.S., Harris, K.S., and MacNeilage, P.F. (1963). A motor theory of speech perception. *Proceedings of the Stockholm seminar*. Stockholm: Royal Institute of Technology.

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-367.

Liberman, A.M., Harris, K.S., Kinney, J.A., and Lane, H.L. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, **61**, 379-388.

Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Mattingly, I., Liberman, A.M., Syrdal, A.K., and Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131-157.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A.M., Jenkins, J.J., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, **18**, 331-340.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D.B., Aslin, R.N., Perey, A.J., and Hennessy, B.L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception & Performance*, **8**, 297-314.

Pisoni, D.B., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.

Pisoni, D.B. and Lazarus, J.H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

Pisoni, D.B. and Tash, J.B. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, **15**, 285-290.

Schwab, E.C. (1981). Auditory and phonetic processing for tone analogs of speech. Unpublished Doctoral Dissertation, Department of Psychology, SUNY at Buffalo.

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, **68**, 836-842.

Strange, W. and Jenkins, J.J. (1978). Role of linguistic experience in the perception of speech. In R.D. Walk and H.L. Pick (Eds.) *Perception & experience*. NY: Plenum. Pp. 125-169.

Tomiak, G.R., Mullennix, J.W., and Sawusch, J.R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America*, **81**, 755-764.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Comprehension of Synthetic Speech Produced by Rule[1]

James V. Ralston, David B. Pisoni, and John W. Mullennix

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This is a draft of a chapter to appear in R. Bennett, A. Syrdal, and S. Greenspan (Eds.), *Behavioral Aspects of Speech Technology: Theory and Applications.* New York: Elsevier (In Press). This research was supported, in part, by NSF Research Grant IRI-86-17847.

Comprehension of Synthetic Speech Produced by Rule

Introduction

Most perceptual evaluations of synthetic speech to date have utilized restricted sets of stimuli presented in isolation for identification and discrimination judgments (see Logan, Greene & Pisoni, 1989). However, because comprehension of speech engages mechanisms beyond those mediating phonemic and lexical recognition, it is likely that the processes involved in comprehension of fluent connected speech cannot be adequately assessed from intelligibility measures alone. Unfortunately, there have been very few studies assessing comprehension of passages of fluent connected speech reported in the literature. In this chapter, we first discuss several preliminaries related to assessing the comprehension of synthetic speech, including the nature of voice output devices, behavioral evaluations of segmental intelligibility, the process and measurement of comprehension, and the role of attention. Next, we review and evaluate the existing body of research devoted specifically to assessing the comprehension of synthetic speech, including recent findings from our own laboratory. Finally, we discuss the major issues that need to be addressed in the future and review some promising techniques that can be applied to these problems.

Voice Output Devices

Our primary interest in this chapter is with rule-based speech synthesis systems, in particular, text-to-speech (TTS) systems. Most consumers are more familiar with systems using stored speech - that is, naturally produced speech waveforms that have been digitally recorded and processed, so they can be played back to listeners at a later time. The chief advantages of stored speech are its high intelligibility and natural sounding speech quality. However, digitally-stored waveforms require large amounts of storage capacity which depends on such factors as sampling rate, amplitude resolution, size of the message set, and length of the utterances. Therefore, digitally encoded speech is more appropriate for applications that require relatively short utterances in a fixed repertoire, such as menus, warning messages, and command systems. Some coding methods (e.g., LPC, CVSD, TDHS) reduce storage requirements, but the storage constraints may still be fairly substantial with long messages or a large corpus of utterances. If the output system is designed for the production of a large or flexible set of messages, stored speech may become unwieldy and unfeasible (Allen, Klatt & Hunnicutt, 1987; Klatt, 1987). In the case of unlimited and unrestricted voice output, the use of stored speech simply becomes impractical. In addition, if one desires to alter the content of the stored messages, the entire recording process must be repeated for each new message. To overcome these problems, it has become common to use synthetic speech produced automatically by rule using a TTS system (Allen, Klatt & Hunnicutt, 1987; Klatt, 1987).

TTS systems are devices that take input text, typically ASCII characters, and transform them into a speech waveform (see Klatt, 1987). In more sophisticated TTS systems such as DECtalk, several linguistic rule-based modules are used to process the phonemic, morphologic, lexical and syntactic aspects of the input string to derive a representation that can be used to produce the final speech waveform (see Allen, Klatt & Hunnicutt, 1987). Although less natural sounding and somewhat less intelligible than natural speech, these systems can produce an unlimited number of messages without recording or storage constraints. Thus, text-to-speech systems are by their design more flexible than waveform coding systems and are ideally suited for applications such as reading machines, computer-assisted learning devices, data-base query systems, and voice mail, all of which must produce unrestricted connected discourse (Allen, 1981).

Behavioral Evaluation

Because of the large number of different TTS systems currently available, it is important to objectively evaluate and compare speech quality using reliable experimental techniques. The most valid method is assessment of human preference and performance in situations resembling actual applications environments. However, this is often not feasible and laboratory studies have had to serve as the benchmark for assessing and comparing different systems. Five major factors have been shown to influence listeners' performance in laboratory situations (Pisoni, Nusbaum & Greene, 1985). These are: (1) the quality of the speech signal, (2) the size and complexity of the message set, (3) short-term memory (STM) capacity of the listener, (4) the complexity of the listening task or other concurrent tasks, and (5) the listener's training state.

Generally, performance is enhanced with higher-quality synthetic speech signals that are closely modeled after natural speech. However, many current speech synthesizers produce impoverished and inappropriate acoustic cues to signal phonetic distinctions. The size of the message set affects listener expectations and indirectly affects performance. Previous research has shown that listeners display higher levels of performance with signals drawn from smaller message sets. STM is one of the most important structural limitations on human performance. STM functions as a general purpose mental workspace with limited processing resources (Baddeley & Hitch, 1974; Klatsky, 1980). Consequently, individuals have a limited ability to encode and process the multitude of sensory inputs impinging on the senses at any given time. The proportion of this limited capacity that different mental tasks require has been shown to be related to the number and complexity of their component subprocesses. Finally, the experience of a listener with a given task can have profound effects on performance with even poor quality synthetic speech (Schwab, Nusbaum & Pisoni, 1985). Perceptual learning allows the listener to adopt a processing strategy that optimizes performance in a given task. The role of each of these factors on the comprehension of synthetic speech will be addressed in the sections below.

Intelligibility

As a first approximation, intelligibility is often assumed to reflect signal quality. Intelligibility measures listeners' ability to recognize different phonemes or words when they are presented in isolation. As such, speech intelligibility provides an index of the lower bounds of perceptual performance for a given transmission device when no higher-level linguistic context is provided (Schmidt-Nielson, this volume). Standardized guidelines have been developed to measure speech intelligibility (ANSI, 1969). However, no standards have been developed to date to measure intelligibility or comprehension of synthetic speech.

Several recent studies have compared the segmental intelligibility of synthetic and natural speech (Hoover et al., 1987; Logan, Greene & Pisoni, 1989; Nusbaum, Pisoni & Dedina, 1984; Nye & Gaitenby, 1973; Pisoni & Hunnicutt, 1980; Pratt, 1987). A popular intelligibility test is the Modified Rhyme Test (MRT), in which a list of 300 monosyllabic CVC words are presented to naive listeners in a forced-choice format (House, Williams, Hecker & Kryter, 1965). Subjects respond by choosing one of six alternative words on each trial. Overall, these studies have found lower intelligibility and different patterns of perceptual errors for synthetic speech compared to natural speech.

In one study carried out in our laboratory, Logan, et al. (1989) obtained MRT scores for synthetic speech produced by ten TTS systems. The results of this study are presented in Figure 1 along with control data from an adult male talker. Performance for different synthesizers varied widely from nearly perfect scores for natural speech to about 90-95% correct for "high-end" synthesizers and 60-70% correct for "low-end" synthesizers. Although some of the segmental confusions for synthetic speech were similar to those observed for natural speech, there were more errors overall for fricative and nasal phonemes. For some systems, the errors appeared to be quite unique to the particular synthesis techniques used.

Insert Figure 1 about here

The same systems and the natural control stimuli were also tested using an open response format test. The original forced-choice form of the MRT provides a closed set of six alternative responses (House et al., 1965). Subjects in the open response condition were required to write their responses to each word on blank lines in a test booklet. Figure 2 shows error data for open and closed formats for the same systems. Error rates for the open response condition were roughly double those observed for the closed response condition across all systems. In addition, the increase in errors from closed to open response set was larger for synthesizers with the greatest closed response errors.

A closer analysis of the pattern of errors indicated a greater diversity of errors in the open response condition. However, the rank order of the errors was the same for the two response

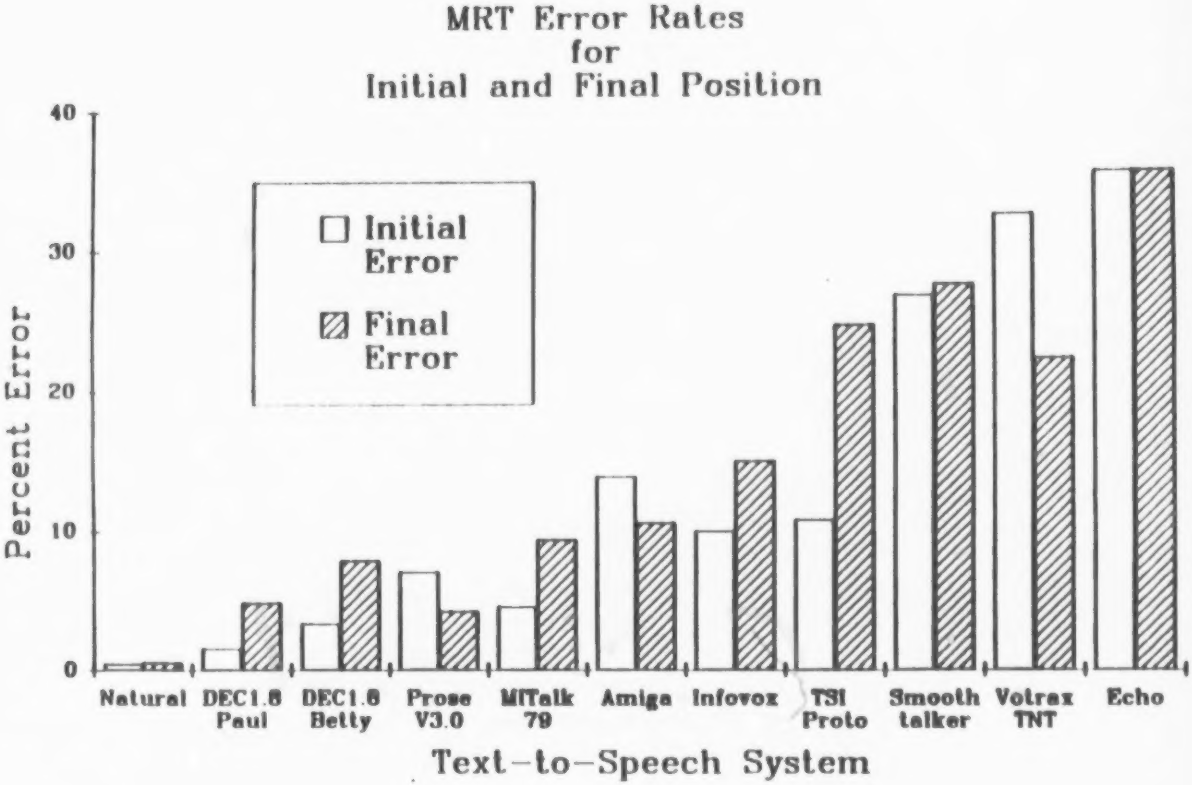


Figure 1. Error rates (in percent) for several TTS systems tested with a closed-format MRT. Open bars represent error rates for syllable-initial contrasts and striped bars represent error rates for syllable-final contrasts (from Logan et al., 1989)

conditions. These results are important because they illustrate the effect of message set size on listener performance. Although the stimuli were the same in both conditions, from the listener's perspective the potential set of stimuli on each trial was much larger in the open response condition. Thus, in a listening task, performance decreases as the size of the message set increases (Pollack & Decker, 1958).

Insert Figure 2 about here

Although differences in segmental structure are generally believed to account for most performance differences between different synthesis systems, the suprasegmental structure of an utterance (i.e., duration, stress pattern and intonation contour) also influences speech intelligibility. Suprasegmental effects operate at two levels. First, inappropriate suprasegmental information often affects some phonetic judgments (Haggard, Ambler & Callow, 1970). Second, other evidence suggests that suprasegmental information provides syntactic and semantic cues to clausal structure and directs attention to certain portions of the speech signal, particularly stressed syllables (Cutler, 1976; Shields, McHugh & Martin, 1974; Wingfield & Klein, 1971). "Low-end" synthesizers typically concatenate phoneme control codes without smoothing, and often do not automatically encode suprasegmental information. For example, the durations of words produced in sentential context by the Votrax Type-N-Talk synthesizer are identical to those produced in isolation (Greenspan, Nusbaum & Pisoni, 1988). On the other hand, systems such as Prose 2000, Infovox, and DECtalk encode substantial prosodic information which is governed by a large number of linguistic rules that consider morphemic, lexical, and syntactic structure of an input text. Thus, differences in intelligibility between various TTS devices are due not only to differences in segmental cues to phonemes but also to differences in suprasegmental information as well.

Comprehension

Relatively few studies have examined the comprehension of synthetic speech, particularly long passages of fluent connected speech. Whereas intelligibility measures assess the perception of individual spoken segments or words, comprehension measures assess a listener's "understanding" of the spoken message, not just the recognition of specific words in the sentence. At the present time, there are no standardized methods for measuring the comprehension of synthetic speech. Webster's New World Dictionary (2nd Ed.) defines comprehension as "the act of grasping with the mind, the capacity for understanding ideas or facts, or the knowledge that results from these processes" (Guralnik, 1986). Psycholinguists have described comprehension as a process by which a listener constructs a coherent mental representation of the propositional information expressed by a passage and relates

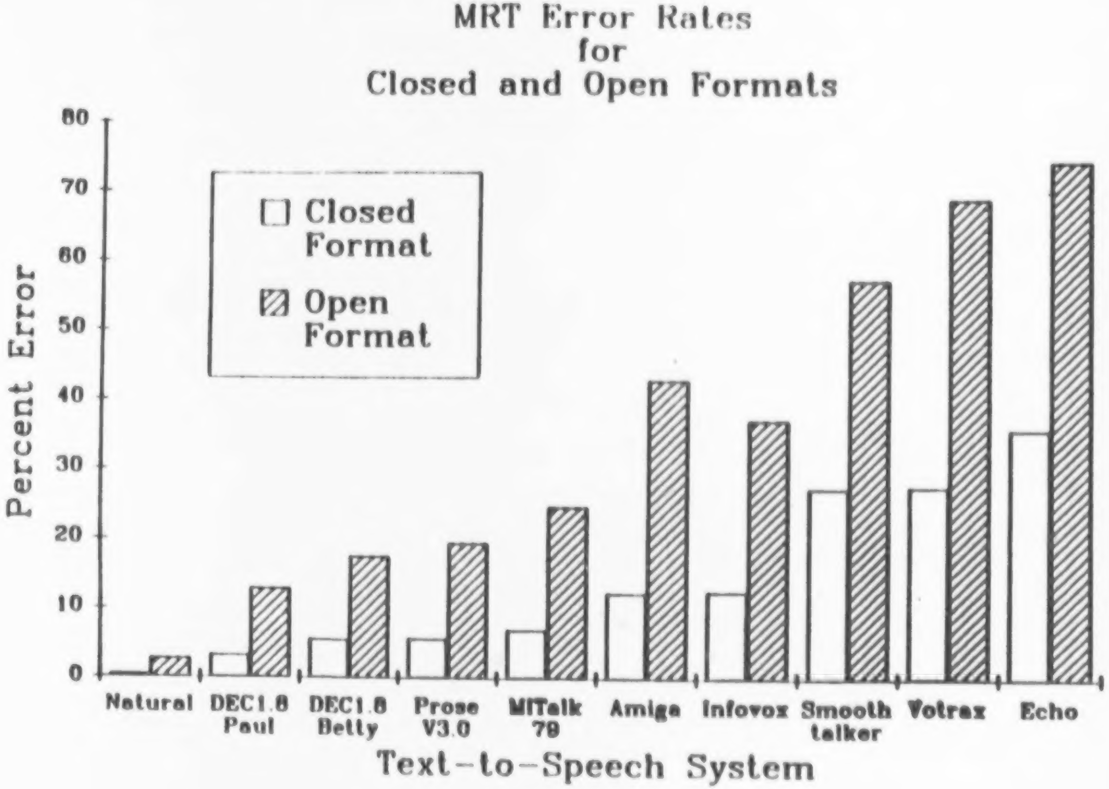


Figure 2. Error rates (in percent) for several TTS systems tested with both closed- and open-format MRT. Open bars represent error rates for the closed-response format and striped bars represent error rates for the open-response format (from Logan et al., 1989)

this structure to other previously or currently available information in memory (Kintsch & van Dijk, 1978). Although human listeners primarily remember the meaning of a passage, evidence also indicates that information about the surface structure of an utterance may also be retained for some time, particularly when task demands favor it (Anderson, 1974; Sachs, 1967).

In spite of the accumulated empirical knowledge, very few coherent theoretical models of comprehension have been developed (Pisoni, Manous & Dedina, 1987). A number of models address relatively restricted domains within the comprehension process. For example, most psychological models of speech perception have emphasized peripheral auditory processing and recognition of linguistic units such as phonemes or words (McClelland & Elman, 1986; Klatt, 1979, 1988; Pisoni & Sawusch, 1975). At the other extreme, Kintsch and Van Dijk (1978) have described a model of reading comprehension that takes a list of meaningful propositions as its input and constructs a coherent information structure that can be used in recall and question-answering tasks. The inputs to this system are presumably produced by syntactic and semantic processes, neither of which are specified in the model. However, there are no integrated models that include all the processing stages that are likely to be involved in comprehension. Despite this problem, it is possible to study comprehension with a variety of techniques.

In the sections below, we make a number of general assumptions about comprehension. First, comprehension is not a monolithic process, but the product of multiple, continuously-interacting processes (Cooper, Shankweiler & Studdert-Kennedy, 1967; Marslen-Wilson & Tyler, 1980). Among the major processes are peripheral and central auditory coding as well as processes related to phonetic, phonological, lexical, prosodic, semantic, syntactic, inferential, and pragmatic information (Pisoni & Luce, 1987). Second, comprehension is a nontrivial mapping of surface structure (the exact sequence of words) onto a more abstract semantic representation (the meaning of the words). Comprehension processes decode the surface structure of sentences, producing propositions representing clausal information. The propositions are connected to one another and to other propositions concerning world knowledge to represent the meaning of an utterance. Third, STM capacity is used by comprehension processes (Baddeley & Hitch, 1974; Kintsch & Van Dijk, 1978). Because of the central role of STM in perception and comprehension, we review briefly several of the major findings in the literature that are relevant to the comprehension of synthetic speech. Then we return to methods of measuring comprehension.

Attention and Processing Resources

Limited Capacity. One of the major structural constraints on human cognitive performance is the attentional capacity of STM in selecting and processing information (Newell & Simon, 1972; Shiffrin, 1976). STM may be conceived of as a "mental workbench," with a limited amount of available space and energy (Baddeley & Hitch, 1974). The energy constraint is often referred to as a limited pool of resources that may be allocated for various tasks (Baddeley & Hitch, 1974; Navon, 1984; Wickens, 1987). Cognitive processes, particularly those which require conscious effort or control, expend certain amounts of the resources (Schneider & Shiffrin, 1977). The demands of various processes may not be apparent when they run in isolation. However, if two or more execute concurrently, and if they place demands on the same limited resources, then performance on any or all may suffer (Navon & Gopher, 1979; Norman & Bobrow, 1975). Processes run simultaneously in most natural situations. For example, an individual driving an automobile can also talk and manipulate other controls at the same time. If the demands of talking or manipulating controls is too great, driving performance may suffer, resulting possibly in more driving errors. It is important to determine the relative resource requirements for various processes. For example, one of our research interests has been focused on whether the perception of synthetic speech demands greater resources than natural speech (Luce, Feustal & Pisoni, 1983). If this is the case, there may be reason to restrict the use of TTS systems to applications where the cognitive load on the listener is relatively low. There is also evidence suggesting the existence of multiple resource pools, each available to different types of processes, such as vision, audition and motor control (Brooks, 1968; Klatt & Netick, 1988; Navon & Gopher, 1979; Shiffrin, 1987; Wickens, 1987). If the resource requirements of separate processes do not exceed the limited capacity of the reservoirs they draw from, and if the processes are not incompatible or interfering, then no performance decrements should be observed when the processes are executed simultaneously (Shiffrin, 1987).

Attention and Synthetic Speech. Several experiments have tested the proposition that the perception of synthetic speech requires greater mental effort or attention. This line of research followed from earlier research with lists of spoken digits (Dallett, 1964; Rabbit, 1966). These studies demonstrated that noise-degraded digits were more poorly remembered in a recall task than undegraded digits, even though both types of stimuli were correctly identified in a labeling task with no memory constraints.

By extension, if the perception of synthetic speech requires more processing resources than natural speech, then one should be able to measure these additional demands with appropriate experimental techniques. If there are differing attentional demands for synthetic speech compared to natural speech, we would expect a differential decrease in performance as the difficulty of a perceptual task is increased or as a second task is added which makes use of the same resources (Baddeley & Hitch, 1974; Luce, Feustel & Pisoni, 1983).

In an important experiment carried out in our laboratory, Luce et al. (1983) tested this hypothesis. In one of their studies, subjects recalled ten word lists produced either by MITalk or by an adult male talker. Before each list of words was presented to subjects, either zero, three, or six digits were displayed in sequence on a video monitor. Subjects were required to first recall the visually presented digits in the order they were displayed and then recall the spoken words in any order. Analysis of the digit recall data indicated that not only were there more errors overall when subjects were presented with synthetic words, but there was a significant interaction between voice and digit preload. More specifically, the decrement in performance for recall of the synthetic speech increased with increasing digit load.

Insert Figure 3 about here

In another experiment, subjects were required to recall in serial order lists of natural words or lists of synthetic words. Results from this study are presented in Figure 3. In addition to finding that synthetic words were recalled more poorly overall than natural words, Luce et al. also found a significant interaction between voice and serial position. That is, the difference in ordered recall between natural and synthetic words was greatest for words from early positions in the lists. Because words from early portions of the list are assumed to be recalled from long-term memory (LTM), the results suggested that increased encoding demands for synthetic speech left fewer resources for the transfer of items into long-term memory (Murdock, 1962; Waugh & Norman, 1965).

In a recent follow-up study, Lee and Nusbaum (1989) studied changes in the processing demands of synthetic speech as a function of practice. Subjects were tested using a target monitoring task before and after three days of identification training with phonetically balanced words produced by a Votrax synthesizer. During pre- and post-training tests, subjects monitored a list of phonetically balanced words for a particular target. Subjects studied two or five visually displayed numbers before the list of words was presented, and recalled them after hearing the list. Although there was an effect of digit load on recall accuracy, there was no effect of training. Word monitoring accuracy improved after training and was higher when there was only a two-digit load, a finding that was consistent with the digit recall data. Monitoring latency decreased after training, and was shorter with only a two-digit load. There was a significant interaction between the effects of training and load on monitoring performance. The effects of training were much larger for the two-digit load condition. Lee and Nusbaum concluded that with training their subjects diverted "spare" resources (the total amount available minus that for the two digit recall task) to acoustic-phonetic processing, which produced shorter monitoring latencies.

Lee and Nusbaum also carried out another experiment to test whether the increased performance observed in the first experiment was due to changes in intelligibility. The authors argued that if the effects of intelligibility and digit load interacted, as training and digit load did in the first experiment, then the training effects seen in the latency data could be accounted for in terms of concurrent changes in intelligibility. The same monitoring and digit recall tasks were employed to determine whether the more intelligible synthetic speech had lower processing costs. Word lists in this experiment were produced by both a high-intelligibility Speech Plus CallText 5000 synthesizer and a low-intelligibility Votrax

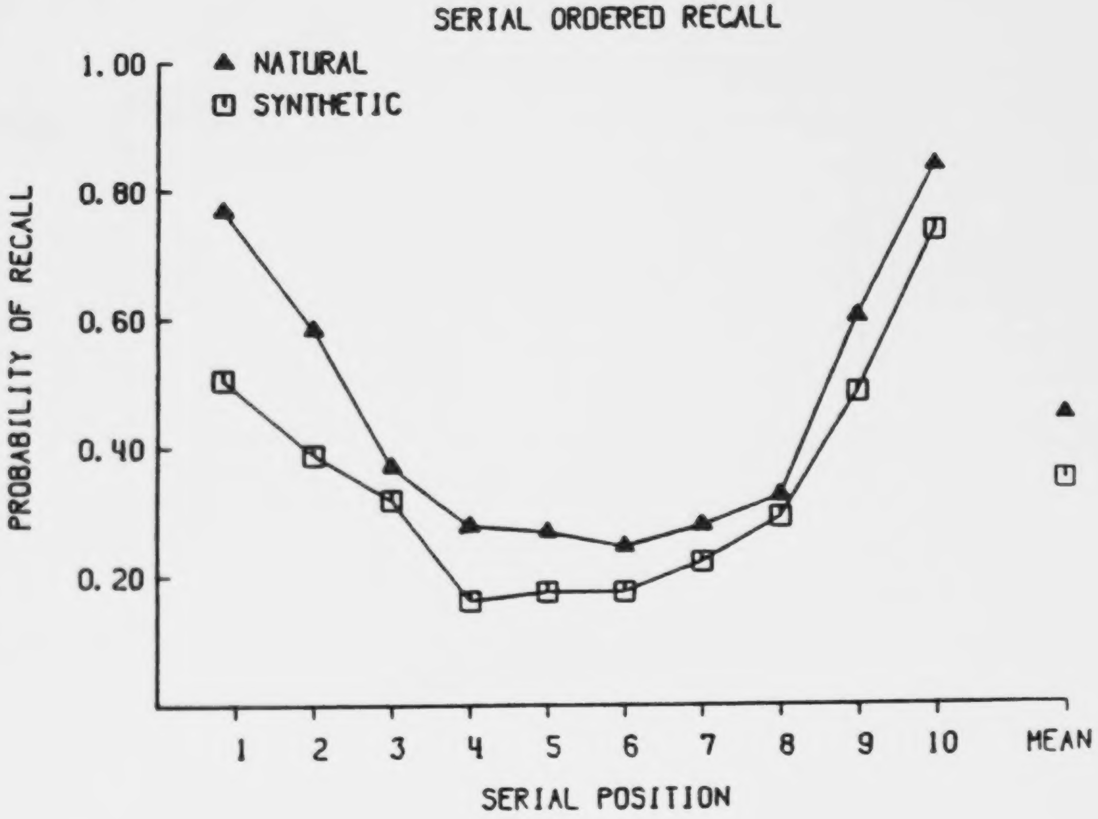


Figure 3. Probability of correct serial recall for natural and synthetic speech (Votrax) word lists. Serial position refers to the presentation order of items the memory set. Triangles represent recall rate for natural speech and squares represent the recall rate for synthetic speech (from Luce et al.,1983).

synthesizer. Subjects saw either two or five numbers in sequence, heard the word lists and then recalled the number sequences. The latency data exhibited significant effects of digit load and intelligibility. CallText words were detected faster than the Votrax words. Words were also detected faster when there was only a two-digit load. However, intelligibility and digit load did not interact. Lee and Nusbaum argued that the interactive effects of voice and load observed in the first experiment were not due to changes in intelligibility per se as a function of training, but instead reflected listeners re-allocation of their available processing resources.

In summary, the available evidence indicates that the perception of synthetic speech requires greater processing resources than natural speech. Because the stimuli in these experiments were simple word lists, the increased demands appear to arise during the acoustic-phonetic analysis of the input signal. In addition, more recent data indicate that given appropriate training, subjects may devote more effort to encoding processes, thereby increasing performance. The following section reviews empirical data suggesting that other comprehension processes may also make use of the same central resources of working memory.

Attention and Comprehension. Several studies have examined the capacity demands of sentence and discourse comprehension. All have demonstrated that comprehension requires STM capacity. In one experiment, Baddeley and Hitch (1974) presented three types of passages to subjects, one sentence at a time, and later administered a Cloze comprehension test (Taylor, 1953). One half of the subjects were required to remember six different digits during each sentence. After the completion of the sentence, subjects attempted to recall the digits in ordered sequence. Baddeley and Hitch found that both comprehension and ordered digit recall were worse compared to a control condition in which subjects that recalled the digits immediately after their presentation. Two other experiments replicated the major finding with different comprehension tests, memory load techniques, and text variables. Based on their results, Baddeley and Hitch concluded that comprehension requires STM processing capacity. Consequently, comprehension performance may suffer if there are other concurrent tasks that also compete for the same STM resources.

Other studies have examined the capacity requirements of some of the subprocesses involved in comprehension. Britton, Holdredge, Westbrook and Curry (1979) asked subjects to read passages with or without titles while they listened for randomly-occurring auditory clicks. Subjects performed more poorly on the click detection task when titles were provided, while comprehension performance also increased. The authors concluded that readers consumed STM capacity while performing the inferences made possible by the titles' context. Therefore, it is likely that some portion of the attentional demands observed by Baddeley and Hitch may have been used for deriving inferences in comprehension.

Other research examining a number of word-level and sentence-level linguistic variables has found local fluctuations in processing load within sentences. Several studies have shown that lexical variables such as a word's frequency of occurrence in the language (Foss, 1969;

Morton & Long, 1976) and ambiguity (Cairns & Kamerman, 1975; Foss, 1970) can influence local processing load. For example, infrequent and ambiguous words lead to longer response latencies in phoneme monitoring tasks (Foss, 1970). Other studies have reported that processing load is influenced by the location of words within a clause or sentence (Abrams & Bever, 1969; Foss, 1969) and the presence of cues to syntactic structure such as relative pronouns (Hakes & Foss, 1970). Processing load is relatively low at the beginning of a clause compared to the end and is reduced in sentences that contain cues to clausal structure (Streeter & Bever, 1975). An important issue, then, is whether or not the increased encoding demands of synthetic speech interact or compete with the attentional demands of other comprehension processes. To summarize, there are a variety of available TTS systems that vary in the sophistication of their synthesis algorithms. These computational differences determine the similarity between the output of TTS devices and natural speech. Several lines of data suggest that the perception of synthetic speech requires more processing resources than natural speech. A quite distinct line of psycholinguistic inquiry has demonstrated that the processes of comprehension also require attentional resources. However, very little research has examined whether the demands imposed by synthetic speech compete for common resources also used in comprehension. The implications of such an interaction are paramount in applications of this technology. Recognizing this, Luce et al. (1983) issued the following caveat:

> "We believe that increased processing demands for the encoding and rehearsal of synthetic speech may place important constraints on the use of various voice-response devices in high information load situations, particularly under conditions requiring differential allocation of attention among several sensory inputs."

Comprehension of connected speech represents a divided attention task in which the listener must devote resources to several cognitive processes that run concurrently in real-time. In the next section, we examine several methods that have been developed to assess comprehension of spoken language. Then we discuss recent studies on the comprehension of synthetic speech.

Measures of Comprehension

Attempts to objectively evaluate comprehension date back at least to the 1800's and are closely related to the development of intelligence testing. Around 1900, the U.S. educational system began developing comprehension tests in order to place students in appropriate classes (Johnston, 1984). This educational tradition has yielded many "standardized" tests that are available in published form. More recently, cognitive psychologists and psycholinguists have developed measurement instruments to assess both the mechanisms and products of comprehension (Levelt, 1978; Carr, 1986). A fundamental distinction has been drawn in comprehension research between successive and simultaneous measures of comprehension. Successive measures of comprehension are those made after the presentation of the linguistic materials. Simultaneous tasks, on the other hand, measure comprehension in real-time as it takes place and usually require subjects to detect some secondary event occurring during the time that the comprehension process takes place (Levelt, 1978).

Successive Measures

Historically, successive measures have been used to study comprehension and language processing much more often than simultaneous measures. Successive measures are appropriate for evaluating what information is abstracted and remembered from text. However, these techniques fail to provide information about whether the observed effects arise during the initial act of comprehension or later during memory storage and retrieval (Levelt, 1978). In addition, successive measures are theoretically less sensitive than simultaneous measures due to memory decay introduced by the retention interval.

Recall. In comprehension tasks that use recall measures, subjects may be asked for verbatim recall, to provide a written summary, or they may be given verbal cues to recall. The linguistic stimuli employed in recall studies have often been connected passages of meaningful text. Bartlett (1932) presented short stories to readers and asked them to reproduce the stories after varying time intervals. He found that subjects not only failed to recall some material from the passages, but they also committed systematic errors suggesting that memory distortions had occurred. Ambiguous portions of the stories were often deleted, new information was added which did not appear in the original passage, and more contemporary terminology was introduced in lieu of antiquated phrases.

More recently, Kintsch and his associates have developed rigorous techniques for objectively scoring the semantic content of recall protocols (Kintsch, 1974; Kintsch & Keenan, 1973; Kintsch, Kozminsky, Streby, McKoon & Keenan, 1975). First, passages are decomposed into their essential propositions or ideas. Later the recalled information is decomposed into its propositional structure and compared to structures obtained from the original passages in order to derive a measure of recall of the propositional content of the passage. Kintsch's research has shown that subjects construct information in recall protocols as well as add metastatements about the text. The probability of these additions increases as a function of the length of the retention interval (Kintsch & Van Dijk, 1978; Levelt, 1978). Thus, the originally comprehended information may be distorted considerably when recall methods are used to measure comprehension.

Recognition. Several types of recognition tests have been devised to assess comprehension, including word and sentence recognition, sentence verification, and multiple choice tasks. In general, research has consistently demonstrated that recall is a less sensitive measure of retention than recognition (Crowder, 1976; Klatsky, 1980).

Word recognition tasks are assumed to assess memory for specific phonological or semantic entities, both of which may be evaluated by the use of rhyming or semantically similar foils (Brunner & Pisoni, 1982). False recognition of rhyming foils indexes memory for the phonological properties of similar target words. False recognition of semantic foils indexes memory for the conceptual properties of similar target words.

Sentence recognition tasks are assumed to assess memory for higher levels of representation, such as propositions derived from text. Appropriate foils may evaluate memory for different properties of sentences. For example, Sachs (1967) presented probe sentences visually after subjects had heard short passages containing embedded "critical" sentences in different locations. The probe sentences were either identical to the critical sentences, identical in meaning but expressed in a different active/passive voice, or completely different in meaning. Although subjects almost always rejected probe sentences expressing different ideas than critical ones, they often incorrectly recognized probe sentences expressing the same ideas as critical sentences in a different voice. This effect was magnified with increasing delays between the critical and probe sentences. These results demonstrate that listeners abstract and remember the meaning of a sentence or passage, but rapidly forget surface structure.

One of the oldest recognition tasks used in comprehension studies is the multiple choice test. After linguistic materials are presented, a question or statement is displayed which either may be completed (in the case of a statement) or answered (in the case of a question) by choosing one of a number of alternatives. This technique has been widely used in reading comprehension research for well over 100 years, and is still employed in most educational settings (Johnston, 1984). One attraction of multiple choice testing is its ease of administration, as compared to other contemporary assessment techniques. However, due to the number of alternatives typically presented, response latencies may be so long and variable as to be unreliable (however, see Nye, Ingeman & Donald, 1975, for a composite reaction time measure).

Sentence Verification. The sentence verification task (SVT) has been used for a variety of purposes in psycholinguistic research. In this task, a test sentence is presented to subjects who are required to judge whether it is "true" or "false." Since the judgments are often trivial (e.g., "A robin is a bird") and error rates are relatively low, the main dependent variable of interest is response latency. This paradigm has been used extensively to study sentence processing (Gough 1965, 1966) and semantic memory (Collins & Quillian, 1969). For example, Gough (1965, 1966) presented active, passive, affirmative, and negative sentences of varying lengths for verification against pictures. The subjects' task was to decide whether the events depicted by the picture and sentence presented on each trial were congruent. Gough found that active sentences were verified faster than passives and that affirmatives were verified quicker than negatives. The results were taken as evidence for a perceptual decoding of surface structure similar to a transformational grammar operating in reverse (Gough, 1966). Clark and Chase (1972) have elaborated processing models of the sentence verification task.

An important assumption of their model is that both linguistic and pictorial information are transformed into a common representational format, which then is used to test for similarity between the two sources of information.

Simultaneous Measures

Although utilized somewhat less often than successive measures, simultaneous, or "on-line" measures provide an experimental methodology to study comprehension processes as they occur in real-time. In contrast to successive measures, simultaneous measures are assumed to be less contaminated by post-perceptual processes that operate at the time of retrieval. However, they are less appropriate for determining what information has been extracted and retained by the subject. Simultaneous tasks usually require subjects to detect some secondary event occurring during the time that the primary comprehension process takes place. Under the assumption that monitoring draws from the same attentional resources as comprehension processes, changes in monitoring error rates and latencies have been used as an index of processing load during comprehension.

The phoneme monitoring task was first developed by Foss (1969; Hakes & Foss, 1970) and has been used extensively ever since to study on-line comprehension processes (Cutler, 1976; Marslen-Wilson & Tyler, 1980; Shields, et al., 1974). Word monitoring tasks have also been used in studies of language processing (Blank, Pisoni & McClaskey, 1981; Brunner & Pisoni, 1982; Marslen-Wilson & Tyler, 1980; Morton & Long, 1976). In both procedures, subjects make a response when a designated target is detected in a sentence or passage. For example, Foss and Lynch (1970) used the phoneme monitoring task to determine whether sentence structure influenced comprehension. They presented either right branching or self-embedded sentences to subjects who were required to monitor for word-initial /b/ phones. Response latencies were longer for the self-embedded sentences, suggesting that they were more difficult to process than the right-branching sentences.

Converging evidence suggests that phoneme detection may occur later than word detection (Blank, et al., 1981; Brunner & Pisoni, 1982; Savin & Bever, 1970). It also appears that phoneme detection is influenced significantly by the properties of the target-bearing word (Morton & Long, 1976). Word monitoring tasks have been modified to assess phonological or semantic representations. Marslen-Wilson and Tyler (1980) had subjects monitor for words that were either identical to a pre-passage cue word, rhymed with the cue word, or were drawn from a semantic category designated by the cue word. The results showed that rhyme monitoring latencies were about 150 ms longer than identity monitoring latencies. The speed of the detection responses suggested that the words were recognized before the physical end of the presented target words. The category responses were about 200 ms longer than the identity responses and exhibited a strong influence of sentential context. The finding that rhyme and category responses were slower than the identity responses is consistent with the assumption that these tasks involve additional processing operations above and beyond word recognition. In addition, the category responses were more sensitive to the presence of sentence context, suggesting changes in the speed of access to semantic information.

Comprehension of Synthetic Speech

Successive Measures

Most studies examining comprehension of synthetic speech have employed successive measures. Only one study used simultaneous measures. The following section reviews studies utilizing sentence-length and passage-length materials. The next section describes a recent study using both successive and simultaneous measures.

Sentence Understanding. Transcription and verification tasks have been used to study the comprehension of natural and synthetic speech using isolated sentences. The transcription tasks were not originally designed to measure comprehension, but instead were intended as easily administered tests of speech intelligibility with open classes of responses (Egan, 1948; Nye & Gaitenby, 1974). However, a recent study has demonstrated that sentence transcription is also sensitive to manipulations affecting comprehension (Manous, Pisoni, Dedina & Nusbaum, 1985). Two types of sentences have been used in the transcription tasks: the Harvard Psychoacoustic Sentences (Egan, 1948), which are semantically and syntactically well-formed (e.g., "Add salt before you fry the egg."), and Haskins Anomalous Sentences (Nye & Gaitenby, 1974), which are syntactically well-formed but semantically anomalous (e.g., "The yellow dog sang the opera."). Since the semantic and syntactic structure of a sentence is known to exert considerable influence on the intelligibility of the component words (Miller, Heise & Lichten, 1951; Miller & Isard, 1963), any differences in performance between Harvard and Haskins sentences would be due to the compensatory effects of semantic context in listening to synthetic speech.

Insert Figure 4 about here

Pisoni and Hunnicutt (1980) obtained transcription scores for natural and synthetic sentences. The synthetic speech was produced by the MITalk system. Transcription accuracy data is displayed in Figure 4. The performance advantage of natural over synthetic speech was greater for the Haskins sentences compared to the Harvard sentences. This interaction between voice and sentence type suggests that listeners presented with MITalk sentences made substantial use of top-down information provided by comprehension processes.

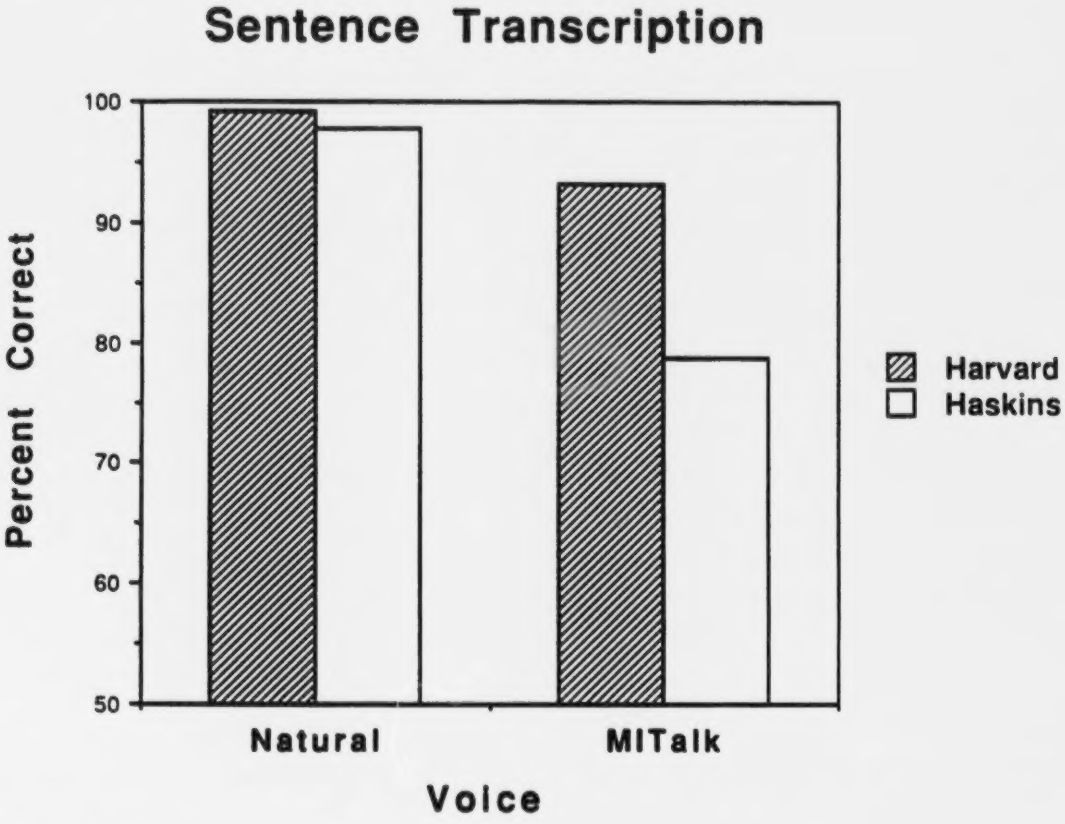


Figure 4. Probability correct transcription as a function of sentence type (Harvard versus Haskins). Open bars represent data for sentences produced by a human voice and the striped bars represent data for sentences produced by the MITalk TTS system. Accuracy is based on the number of individual words correctly transcribed (from Pisoni & Hunnicutt, 1980).

Logan and Pisoni (1986) have also examined differences in transcription accuracy between natural and synthetic speech using the recently developed Phoneme Specific Sentences (Huggins & Nickerson, 1985). The materials for this test were constructed so that different sets of sentences contained a high proportion of different phoneme classes. Three TTS systems (DECtalk, Prose, Infovox) and a natural speech control were tested.

In general, accuracy was relatively low compared to that found in other intelligibility and sentence transcription studies. This difference probably reflects both the stringent scoring criterion (whole sentence accuracy) as well as the infrequent words and sentence structures that were adopted in order to meet the phonemic composition constraints. More importantly, differences in overall accuracy were observed between natural speech and the three types of synthetic speech. This result is consistent with previously obtained segmental intelligibility data. However, Logan & Pisoni failed to find a significant difference between DECtalk and Prose, when earlier intelligibility studies found reliable differences between the two (Logan et al., 1989). Finally, Logan & Pisoni found different patterns of errors for the different types of speech which were consistent with those reported by intelligibility studies (Logan et al., 1989). Thus, the transcription data collected with Phoneme Specific Sentences parallel those collected with isolated syllables or words in intelligibility tests.

The Speech Perception In Noise (SPIN) test manipulates the degree of sentence context within a single test (Kalikow, Stevens & Elliot, 1977). Subjects in this task are presented eight sets of 50 sentences, all masked by a background "babble" noise, and are required to write only the last word of each sentence. In one half of the sentences, the final word is highly predictable on the basis of the preceding words; in the other half, the final word is not predictable. The difference in performance between these two types of context indicates the contribution of higher-order processes to intelligibility of words in sentence context. The SPIN test has been used only once to study synthetic speech in a preliminary investigation by Chial (1984). In this study, overall accuracy was better for a natural speech control than for four low-quality TTS systems, two built around Votrax synthesizers, and two built around Echo synthesizers. In addition, performance was higher for the high predictability sentences than for the low predictability sentences, a finding that is consistent with the results of Pisoni and Hunnicutt (1980).

Insert Figure 5 about here

In several recent studies, transcription responses were obtained from subjects immediately after verification judgments in order to evaluate the relationship between intelligibility and comprehension (Manous, et al., 1985; Pisoni & Dedina, 1986; Pisoni, Manous & Dedina, 1987). Manous et al. (1985) presented sentences produced by two natural talkers and five TTS systems (DECtalk-Paul, DECtalk-Betty, Infovox, Prose, and Votrax Type-n-Talk).

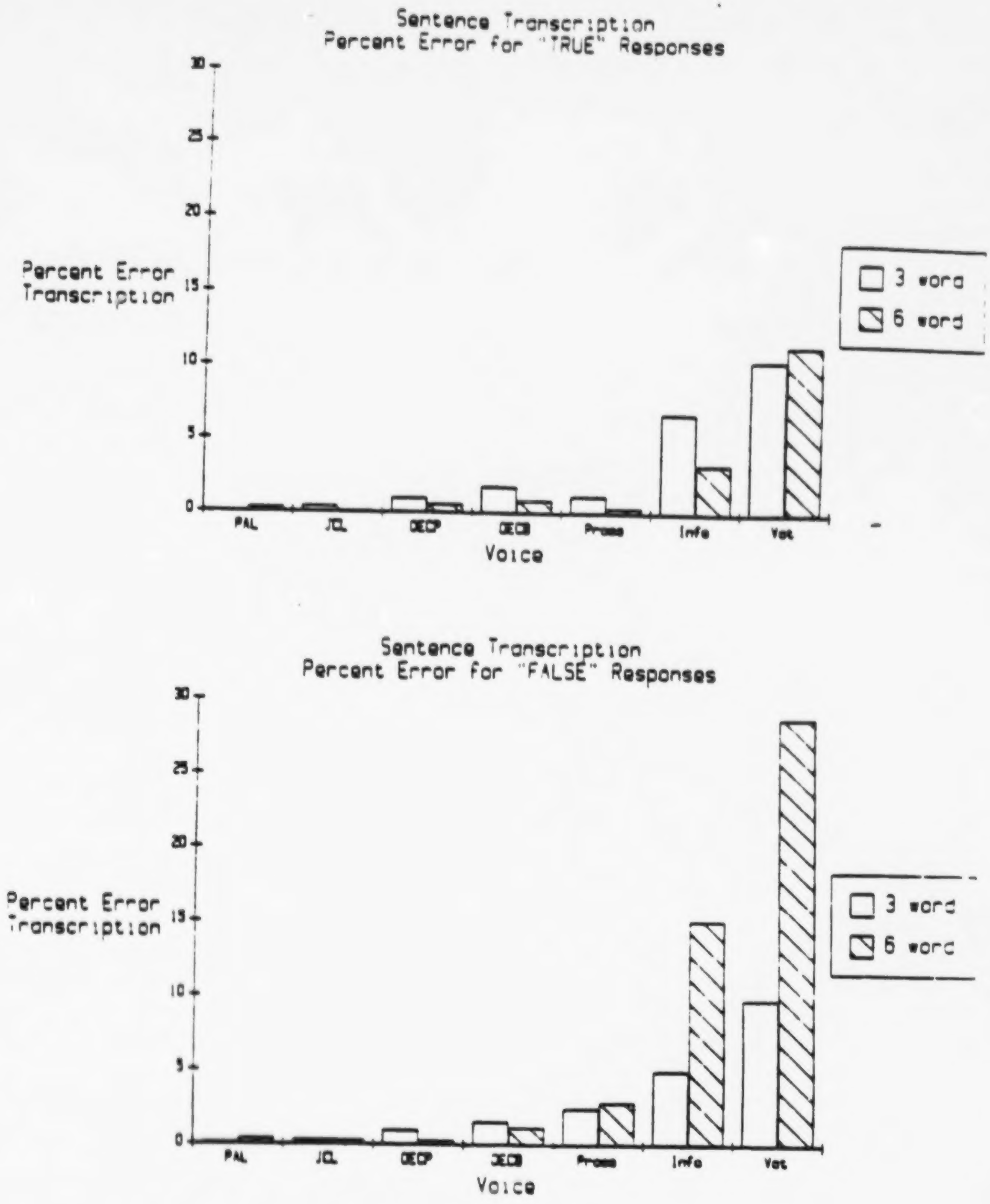


Figure 5. Percentages of errors obtained in a sentence transcription task for "True" repsonses (top panel) and "False" responses (bottom panel) for seven different voices. Open bars represent data for three-word sentences and striped bars represent data for six-word sentences (from Manous et al., 1985).

The sentences were either three or six words in length and expressed either true or false information with respect to general world knowledge.

Figure 5 shows the error rate on the transcription task. Transcription accuracy was generally lowest for Votrax and Infovox synthetic speech and was lower for the six word sentences than the three word sentences. In addition, an interaction was observed between voice and sentence length for the false sentences. That is, errors rates were greater for the longer sentences, but only for sentences produced by the Votrax system. The results are consistent with the assumption that the increased perceptual demands of Votrax and longer sentences tax a common resource pool. As processing demands increase, differences between natural and synthetic speech become larger.

Insert Figures 6 and 7 about here

In addition to transcribing the sentences, subjects in the Manous et al. (1985) study were also required to assess the truth value of the same sentences. Figures 6 and 7 display accuracy and latency data, respectively, from the sentence verification task (SVT). Analysis of variance on both sets of data revealed main effects of voice and interactions between voice and sentence length for false sentences. The subjects who listened to the Votrax sentences were less accurate and slower to respond compared to the subjects who listened to the other speech. In addition, the decreased accuracy associated with the longer sentences was especially marked for Votrax sentences. Post-hoc analysis of the latency data discriminated three groups of voices: natural speech, high-quality synthetic speech (DECtalk and Prose), and moderate-to-poor quality synthetic speech (Infovox and Votrax).

Taken together, the transcription results demonstrate that words in sentences spoken by natural talkers are recognized better than words produced by low-quality synthetic systems. The verification data also demonstrate that sentences produced by natural talkers are also comprehended faster and more accurately than sentences produced by TTS systems. The interactions between voice and sentence length noted by Manous et al. provide additional support for Luce et al.'s (1983) conclusion that encoding synthetic speech incurs greater processing costs, and that these demands may interact with the demands imposed by other task variables, such as comprehension.

Manous et al. (1985) also computed correlations between transcription accuracy, verification accuracy, and verification latency. Considering only the true sentences, transcription accuracy and verification accuracy were highly correlated ($r = -.86$). For false sentences, transcription accuracy and verification latency were also highly correlated variables ($r = +.75$). The authors concluded that both verification and transcription tasks are sensitive and reliable indices of comprehension performance.

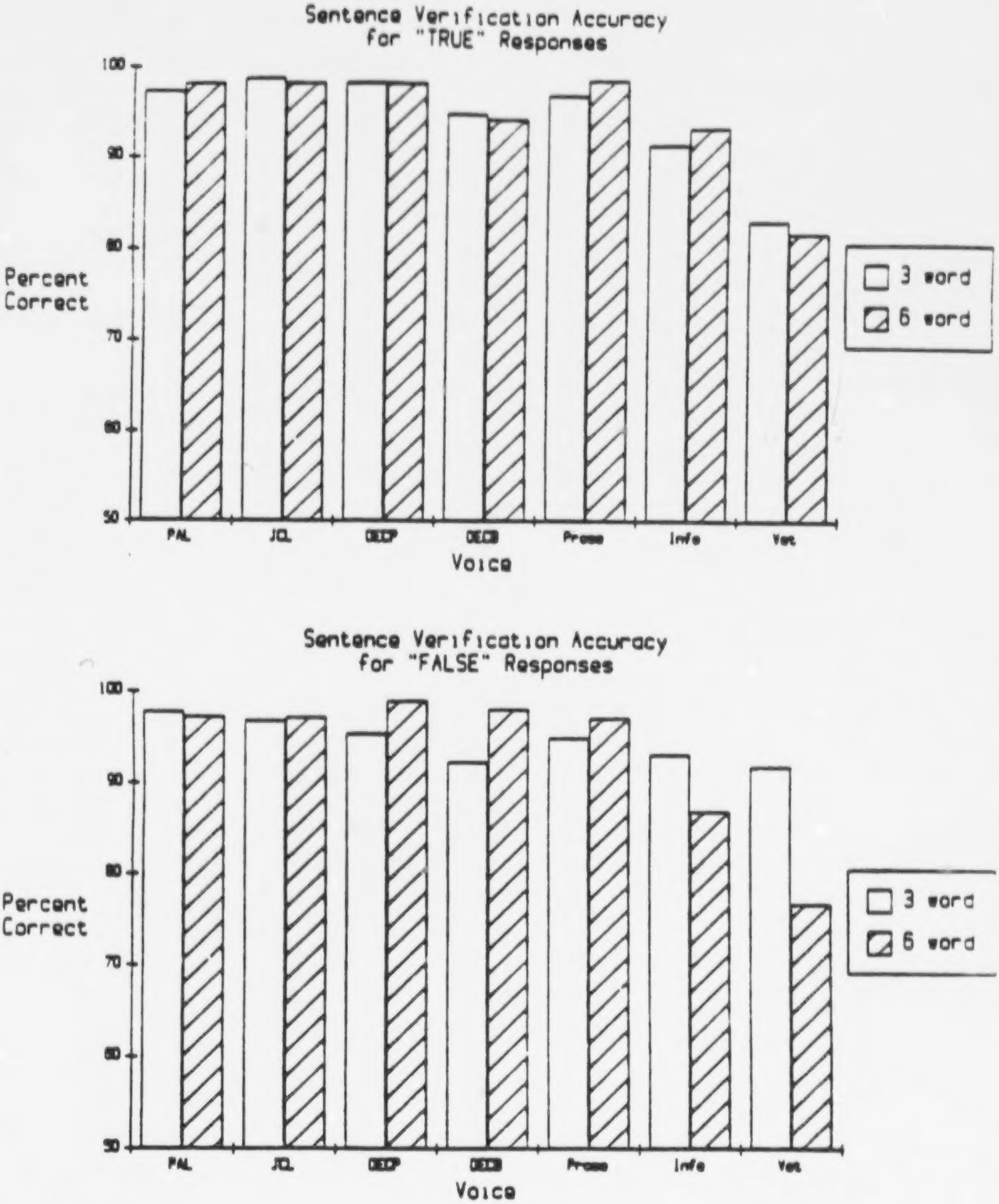


Figure 6. Sentence verification accuracy scores for "True" responses (top panel) and "False" responses (bottom panel) for seven different voices. Open bars represent data for three-word sentences and striped bars represent data for six-word sentences (from Manous et al., 1985)

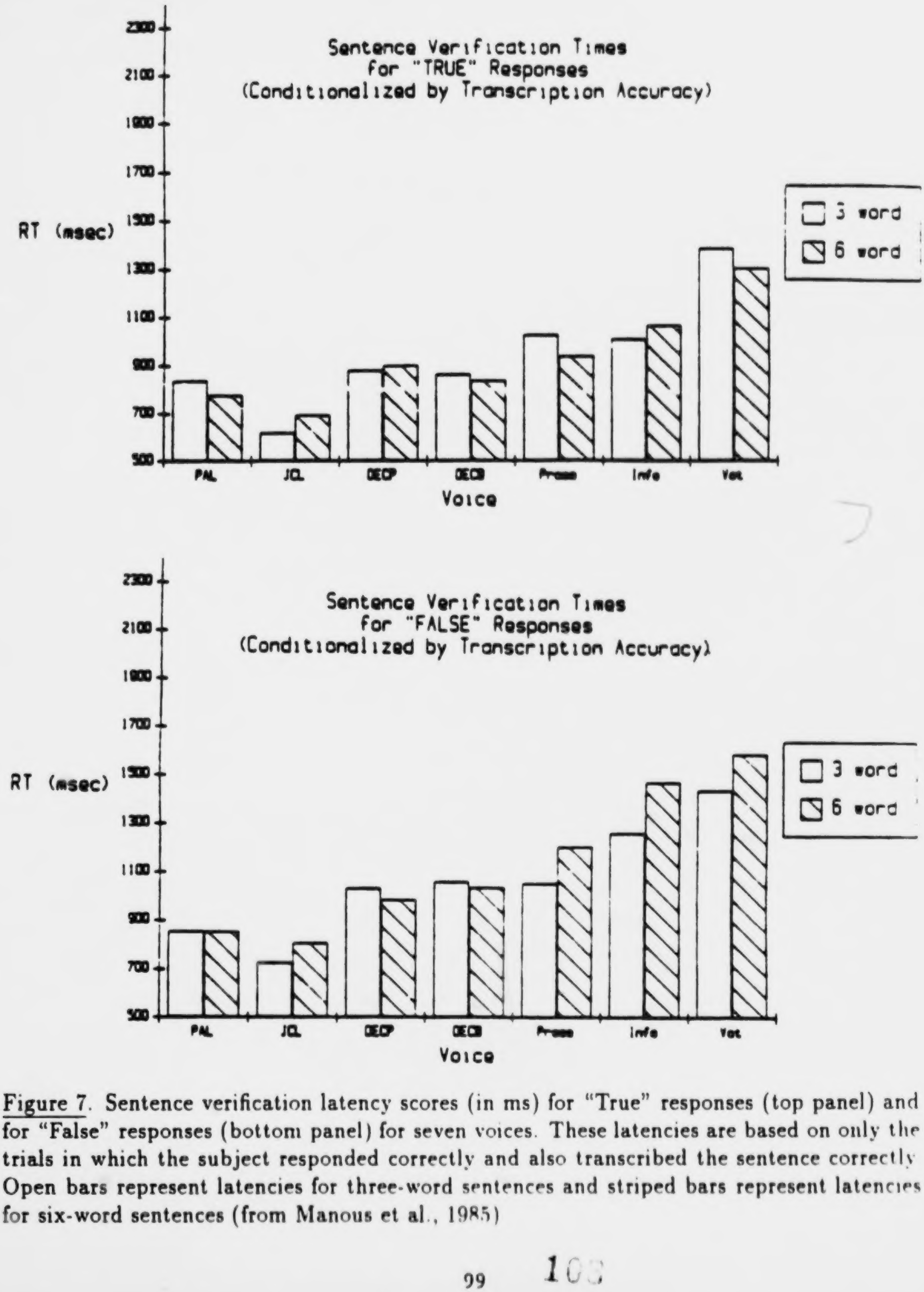


Figure 7. Sentence verification latency scores (in ms) for "True" responses (top panel) and for "False" responses (bottom panel) for seven voices. These latencies are based on only the trials in which the subject responded correctly and also transcribed the sentence correctly. Open bars represent latencies for three-word sentences and striped bars represent latencies for six-word sentences (from Manous et al., 1985)

Insert Table 1 about here

Because the patterns of verification performance observed in the Manous et al. study could have been due to mis-identified words, Pisoni et al. (1987) carried out another study that was designed specifically to dissociate intelligibility from comprehension performance. Natural and synthetic sentences generated by DEC-Talk Paul were initially prescreened using a transcription task. A final set of sentences was selected so that all of them were transcribed nearly perfectly. These sentences contained predicates that were either predictable or not predictable in the sentence frame. Table 1 lists examples of the two different sentence types. After hearing each test sentence, subjects were required to first judge its truth value and then write a transcription in an answer booklet. Verification latency data from this study are presented in Figure 8.

Insert Figure 8 about here

There was no effect of voice, predictability, or length on the verification accuracy data, a result that was consistent with the screening treatment. However, significant effects of all three variables were observed for the verification latencies. The synthetic sentences were verified slower than the natural sentences, even though, according to the transcription scores, the component words were recognized equally well. Longer sentences were verified slower than shorter sentences, and sentences with unpredictable predicates were verified slower than those with highly predictable predicates. No interactions were observed between any of the main variables. The voice effect suggests that synthetic speech, even very high quality speech that can be transcribed accurately, requires more encoding time. The slower response times appear to propagate up the system to higher-level comprehension mechanisms.

Based on the results of Manous et al. using the SVT, Schmidt-Nielson and Kallman (1987) conducted an experiment using the same methodology with digitally encoded speech stimuli. Subjects verified sentences that differed in their truth value and the degree to which their subjects and predicates were associated. Sentences were produced by a male talker and then processed digitally. The experimental sentences presented to subjects were either the original unprocessed versions or LPC-encoded copies. The LPC-encoded versions had either 0%, 2% or 5% bit errors added as noise.

Analysis of accuracy and latency data revealed effects of voice, noise, relatedness, and practice. Accuracy was higher and latency was shorter for high-relatedness sentences, for sentences with less coding noise, and sentences from the second half of the testing session. Using previously obtained DRT intelligibility scores (see Voiers, 1983, for a description of

Table 1

Stimulus materials from Pisoni, Manous & Dedina (1987).

A. Three-word, false, high-predictability sentences

1. Men wear dresses.
2. Circles are square.
3. Sandpaper is smooth.
4. Winter is hot.
5. Screaming is soft.

B. Six-word, true, low-predictability sentences

1. Fish can swim but can't smoke.
2. Smoking is bad for your teeth.
3. Our alphabet has 26 characters.
4. A triangle has only three vertices.
5. Hawaii's a good place to sunbathe.

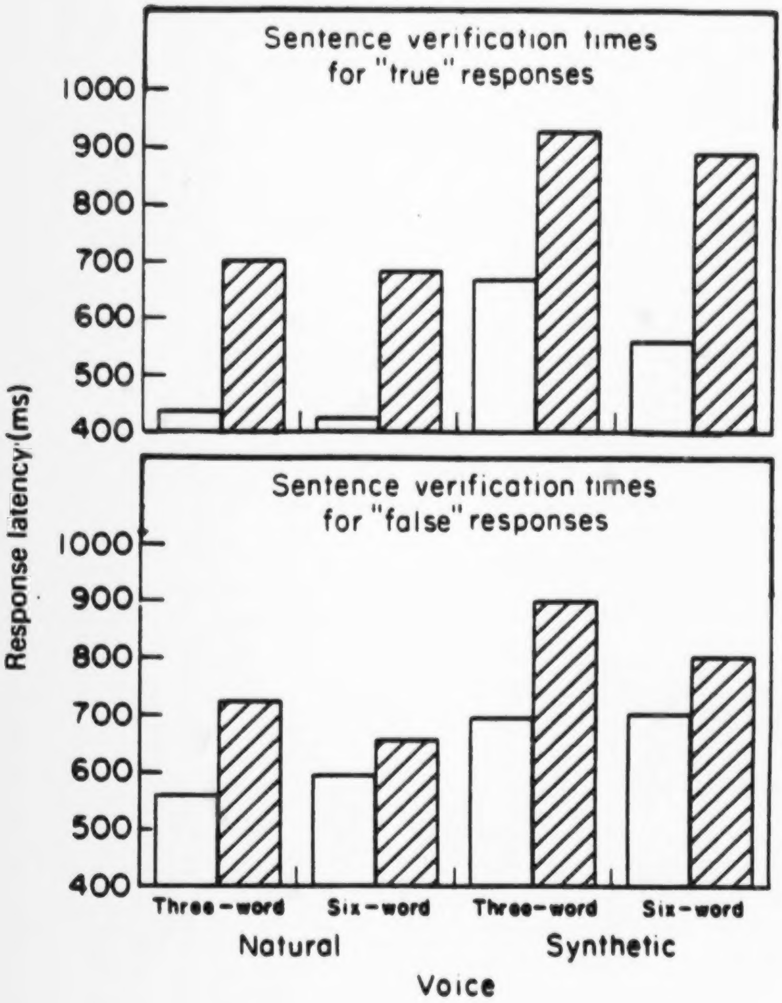


Figure 8. Sentence verification latency (in ms) for "True" responses (top panel) and "False" sentences (bottom panel) for natural and synthetic (DECtalk) speech. Open bars represent latencies for sentences in which the predicate was not predictable from the sentence context, and striped bars represent latencies for sentences in which the predicate was predicatable from the sentence context (from Pisoni et al., 1987).

the technique), Schmidt-Neilson and Kallman estimated that an increase of one percentage point in intelligibility accuracy was related to a decrease of 10-20 ms in verification latency.

The results of this study show that when listeners are presented with digitally-encoded or degraded signals, they make much greater use of contextual information. Pisoni et al. (1987) did not find an interaction between sentence voice and predictability. However, the difference in results may be accounted for by stimulus and subject state variables. For example, Pisoni et al. used highly-intelligible synthetic speech which was not degraded, a different form of sentence predictability, and relatively unpracticed subjects. Schmidt-Nielson and Kallman suggested that when subjects first listen to processed speech, they devote most of their available attention to acoustic-phonetic processing. After a period of familiarization, the acoustic-phonetic processing requires less resources (see Lee & Nusbaum, 1989). These "spare" resources then become available for the comprehension process which generates lexical and phonemic expectations to aid the listener in extracting the message.

Taken together, these recent experiments suggest that sentence transcription and verification are very sensitive measures of comprehension. All of the studies demonstrated differences between natural speech, synthetic speech and processed natural speech. Interestingly, Pisoni et al. (1987) showed that verification latency was sensitive to differences in voice quality even after the segmental intelligibility of the two sources had been equated. These findings suggest that the speed and efficiency of comprehension for different systems may vary substantially for different voice output devices, even if intelligibility is equivalent. However, the locus of this effect is still unknown. It is quite possible that the differences may arise simply from the increased perceptual encoding demands of the synthetic speech. If so, the increased encoding demands may limit higher levels of processing, such as those involved in parsing, semantic activation, or inference-driving. Research is currently underway on these problems and any definitive answers must await the outcome of these studies.

In summary, the interactions observed between voice and the other variables are consistent with a limited capacity resource framework. Manous et al. (1985) found an interaction between voice and length of sentences - performance was particularly poor for long sentences produced by the Votrax synthesizer. This finding suggests that processing synthetic speech and processing longer sentences both impose demands on the same limited pool of processing resources. This view is supported by results from similar studies of processed natural speech which found interactions between voice and sentence length (Pisoni & Dedina, 1986) and voice and subject-predicate relatedness (Schmidt-Nielson & Kallman, 1987).

Comprehension of Fluent Connected Speech. The first study to investigate the comprehension of fluent synthetic speech was carried out by Nye, et al. (1975). The outputs of two synthesizers, the Haskins parallel formant synthesizer and the OVE-III serial formant synthesizer were compared to natural speech. Two passages over 1500 words in length were selected from a published reading test. Each passage was followed by 14 multiple-choice questions. A separate test indicated that subjects could not answer the questions on the basis of world knowledge alone. After each passage was presented, subjects were given as much time as needed to answer as many questions as possible.

The results showed no difference in performance between conditions in terms of accuracy. However, there was a difference in terms of the amount of time needed to answer the questions after replaying portions of the passage. Subjects were nearly 25% slower answering the subset of questions following synthetic passages (mean = 6.27 s) compared to natural passages (mean = 4.51 s). However, since the natural passages were spoken at a faster rate and were therefore physically shorter than the synthetic passages, subjects may have been able to review the natural passages more quickly, thus accounting for the observed latency differences.

In another early study, McHugh (1976) employed eight passages from the Diagnostic Reading Scales, a standardized comprehension test, to study the comprehension of Votrax synthetic speech. Two of the passages were presented as practice and six were presented for the experiment proper. Each of the passages was a short narrative story and each was followed by a series of cued recall questions (e.g., "How much did Bob pay for the plum?"). Seven versions of the paragraphs were presented to subjects, six produced by a Votrax synthesizer and one produced by a human talker. The Votrax versions differed in their stress patterns, which were altered either by hand or by rule. A random stress condition was also included. The stress patterns generated by rule varied in sophistication from relatively crude and mechanical to relatively natural. For example, one algorithm created sentences by alternating stressed and unstressed syllables, while another made use of syntactic information. Voice (talker and stress-algorithm) was varied as a between-subjects factor.

McHugh's results revealed no difference between the different voice conditions - all were comprehended equally well in terms of scores on the recall questions. Data from the two practice passages were analyzed separately. The results showed that the natural passages were comprehended better than some Votrax versions, such as the "untreated" monotone version. However, the natural practice passages were not comprehended any better than the hand-altered Votrax passages with correct English stress patterns. The differences in performance between the practice and test data suggests that even moderate amounts of familiarity and practice are sufficient to allow listeners to quickly learn to process even poor-quality Votrax synthetic speech.

Pisoni and Hunnicutt (1980) reported the results of a study designed to assess the comprehension of MITalk, a forerunner of the DECtalk system. Subjects listened to either DECtalk or natural passages, or, in a control condition, read the same passages. The materials were derived from a variety of published reading comprehension tests. Each passage was followed by a series of multiple-choice questions. Accuracy data from this study are presented in Figure 9. Overall, the reading group performed at a higher level than the MITalk or natural groups. However, when the data were divided into first and second halves of the test session, an interaction between voice and session emerged. Reading and natural speech performance were nearly the same in both halves, but performance for the MITalk group increased dramatically in the second half of testing. As in the McHugh (1976) study, there appeared to be rapid perceptual learning which compensated for early differences in performance.

Insert Figure 9 about here

Jenkins and Franklin (1981) conducted two experiments on the comprehension of synthetic speech. In one experiment, the hand-applied stress and random stress versions of McHugh's Votrax stimuli were presented to subjects for subsequent free recall. Only three of the original passages were tested. No significant difference in performance was observed between the two kinds of speech.

In a second experiment, natural and synthetic speech versions of a passage were presented twice to one group of subjects for an intelligibility test and to a second group of subjects for a comprehension test. The synthetic speech was produced by the OVE speech synthesizer and an experimental TTS system under development at Haskins Laboratories. Two subgroups of synthetic speech listeners were used – those with and without practice. The practiced subjects in this experiment had been tested with twenty passages a week prior to the actual testing. Trial 1 dictation data revealed relatively small but reliable differences in performance between the natural and synthetic conditions. However, Trial 2 data revealed no difference between the natural and practiced synthetic subjects. Therefore, intelligibility differences were eliminated with even modest training. The comprehension test produced comparable results. That is, recall performance on Trial 1 and Trial 2 was better for the natural group compared to the unpracticed synthetic group, but the natural speech group was not significantly different from the practiced synthetic group. Again, initial comprehension differences between natural and synthetic speech decreased with relatively little practice.

Luce (1981) compared recognition memory for different types of information about passages of natural speech or MITalk speech. After each passage was presented, a series of verification sentences was displayed on a video monitor in front of each subject. Different sentences probed the listeners' memory for either lexical information or for different types of propositional information in the text base. Lexical items were verified faster than propositional information, but no differences were observed between natural and synthetic speech. However, verification accuracy was worse for the MITalk passages compared to the natural passages. Finally, Luce observed an interaction between voice and sentence type for the accuracy data. Accuracy was better for MITalk compared to natural speech for the "surface" sentences, which probed lexical memory, but was worse for MITalk compared to natural speech for the propositional sentences. Luce suggested that subjects listening to the MITalk passages allocated a greater proportion of their resources to acoustic-phonetic processing,

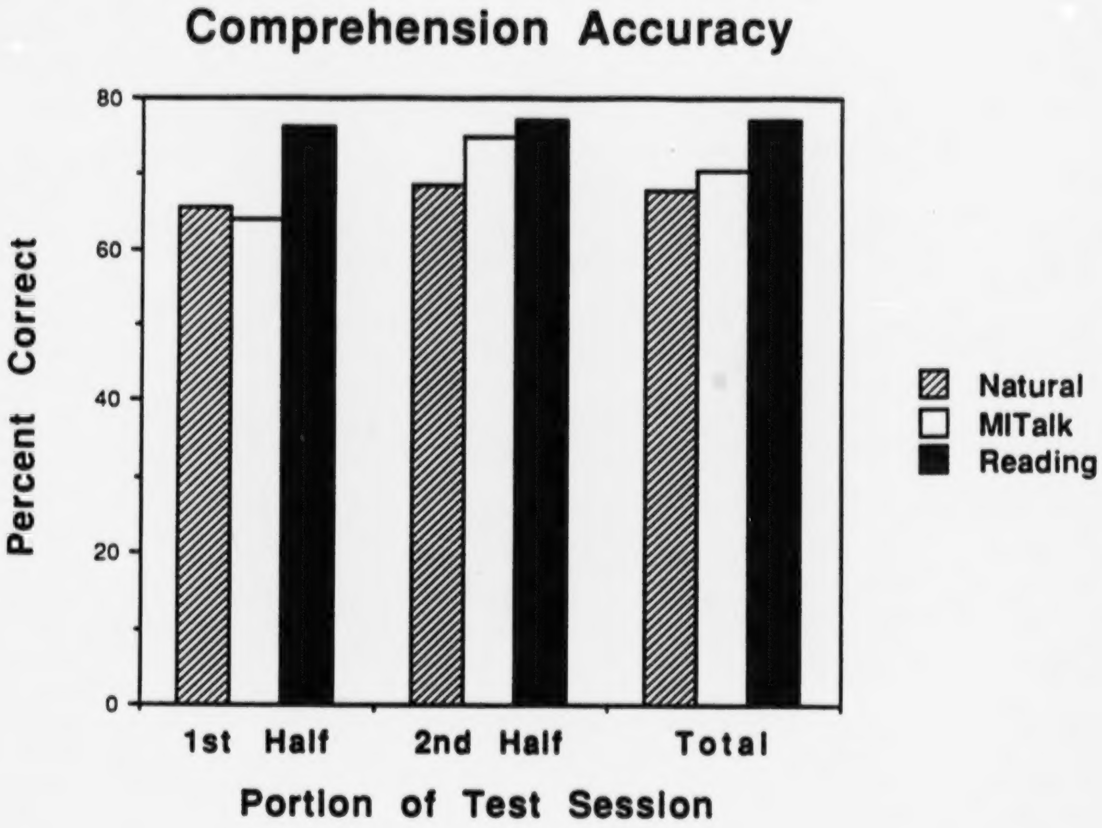


Figure 9. Comprehension accuracy for questions following passages of natural, synthetic (MITalk) speech for the first and last halves of the testing session. Striped bars represent accuracy for questions following natural speech passages, open bars represent accuracy following MITalk passages, and the filled bars represent accuracy in a reading control group (from Pisoni & Hunnicutt, 1980).

which resulted in a more durable memory code for lexical information. Consequently, less resources were available for processing propositional information, a strategy that produced a more fragile propositional memory code.

Hersch and Tartarglia (1983) studied the comprehension of DECtalk Paul and DECtalk Betty at several presentation rates. The data were compared to previous data collected with natural speech by Fairbanks, Guttman & Miron (1957). Subjects were presented short passages over a telephone. After each of the passages, a series of questions was presented that either assessed memory for explicitly stated information or required inferences based on information from the passage. The results indicated that comprehension of the synthetic passages decreased faster as a function of presentation rate than the natural speech (Fairbanks, et al., 1957). Hersch and Tartarglia argued that this finding reflected the increased encoding demands for synthetic speech compared to natural speech. In addition, comprehension accuracy was higher for the male voice, with more practice, and for inferential questions. However, the authors observed that subjects knew the answers to the inferential questions without hearing the passages. While overall accuracy increased across testing blocks, the differences in performance between female and male voices decreased across testing blocks.

As part of a larger study, Schwab, et al. (1985) assessed the comprehension of Votrax passages as a function of training. During the same sessions, subjects were given both intelligibility and comprehension tests. Pre-training and post-training tests consisted of presentations of Votrax passages followed by verification statements assessing different levels of comprehension. On each of the training days, four passages were presented, each followed by a series of multiple choice questions. One group of subjects was trained with Votrax stimuli, another group was trained with natural speech stimuli, and a third group, the control group, was given no training stimuli. All training groups performed better on the comprehension test before training, and no reliable effect of training condition was observed across any of the comprehension measures. In addition, an examination of the training data revealed no difference between the Votrax and natural speech groups, nor a significant increase in performance across training sessions. No interaction was observed between voice and sentence type, as Luce had found earlier. It is possible that subjects in this study reached a training plateau by virtue of the segmental tests by the time of comprehension testing. Alternatively, as Schwab et al. suggested, these tests may have been too sensitive to subject differences, and the difficulty of the various passages may have been confounded with the day of training and/or testing. In any event, it is not clear why the earlier Luce findings were not replicated in this study.

Another investigation of comprehension of synthetic speech was conducted by Moody and Joost (1986). They compared comprehension of passages of natural speech, DECtalk Paul, 9600 bps digitized speech, 2400 bps digitized speech, and a reading control. Passages and multiple choice questions were drawn from study guides for three standardized exams: the GED (a high school equivalence exam), the SAT (an undergraduate entrance exam) and the GRE (a graduate school entrance exam). A fourth factor of question type was analyzed separately. Analysis of the accuracy data indicated that voice, exam type, and length all exerted significant effects on performance. Natural speech was comprehended better than either DECtalk Paul or 2400 bps digitized speech, the SAT was the most difficult type of passage, followed by the GRE, and finally the GED, and the moderate length passages were comprehended better than the long and short passages, which did not differ.

Moody and Joost (1986) also classified the experimental questions by the type of information processing required for their correct solution. Some of this information included use of world knowledge, recognition of information explicitly stated in the passage, and the drawing of inferences from textual information. Based on an analysis of the amount of mental effort required to answer the questions, the questions formed a difficulty gradient from those requiring use of world knowledge to those that required subjects to make difficult inferences.

The results showed that subjects were correct on 100% of the world knowledge questions and only 27% of the low inference questions. Further analyses showed that the advantage of natural speech over synthetic speech became smaller as the difficulty of the questions increased. Comprehension performance was better for natural speech compared to synthetic speech for questions which tested memory for explicitly stated information but not for questions requiring inferences.

The results of this study should be viewed with some caution, however. First, question type was poorly controlled and possibly confounded with passages in this study. Second, text difficulty was also poorly controlled. This was born out by the unexpected rank ordering of perceived difficulty and performance as a function of passage type. Finally, the specific interaction reported between question type and voice appears anomalous. As Pisoni et al. (1987) observed, "We do not know of any current theory of human information processing or language processing that would predict the results observed by Moody & Joost." In retrospect, however, it appears that the effect of voice may have been the most robust result of the study.

The comprehension studies summarized above may be classified into three broad categories. First, two studies reported reliable effects of voice on accuracy, but no training effects (Luce, 1981; Moody & Joost, 1986). Second, three studies reported performance differences between voices when synthetic speech was first encountered during the experiment, but the differences became smaller with even moderate exposure or training (Jenkins & Franklin, 1981; McHugh, 1976; Pisoni & Hunnicutt, 1980). Finally, Schwab et al. (1985) found no comprehension differences between natural and Votrax passages, nor a learning effect on the comprehension of Votrax passages, even after two weeks of training. Taken together, these studies suggest differences in comprehension between natural and synthetic speech, but the results appear to be extremely variable from study to study. Obviously, further research with greater experimental control is necessary to provide more reliable information about the comprehension of synthetic speech.

Simultaneous Measures

One of the first studies to use simultaneous measures to assess comprehension of synthetic speech employed a sentence listening procedure (Mimmack, 1982; Mimmack & Pisoni, 1982) that was analogous to sentence reading procedures (Aaronson & Scarborough, 1976; Cirilo & Foss, 1980). In these experiments, subjects controlled the onset of successively presented sentences produced by a natural talker or a TSI Prose 2000 TTS. The major dependent variable was the latency from the end of each sentence to the subject's response to initiate the following sentence. One half of the subjects were presented sentences in a normal order, while the other half was presented the sentences in a randomized order. Within each condition, half of the subjects ("Comprehension Condition") were required to answer comprehension questions after each passage; the other half ("Recall Condition") were required to recall the passage verbatim. Based on earlier research, both of these variables were shown to strongly influence reading times (Aaronson & Scarborough, 1976; Cirilo & Foss, 1980; Kintsch, Mandel & Kozminsky, 1977).

Insert Figure 10 about here

The results from an initial study using natural speech are displayed in Figure 10. The response latencies of subjects in the recall condition were significantly longer than the latencies in the comprehension condition. In addition, latencies were longer for sentences presented in a scrambled order compared to sentences in the normal order. Both findings were consistent with previous reports on sentence-by-sentence reading times (Aaronson & Scarborough, 1976).

Insert Figure 11 about here

A second study (Mimmack & Pisoni, 1982) used the same methods and design but also included a set of Prose synthetic stimuli. The response times from this experiment are displayed in Figure 11. In the verbatim recall condition, sentence-by-sentence latencies were longer for sentences produced by Prose than those produced by a real talker. The results suggest that the on-line processes used in comprehension of synthetic speech is indeed slower than the processes used in comprehension of natural speech.

Recently, we have combined an on-line word monitoring task with several successive recognition memory tasks to assess processing differences between natural and synthetic

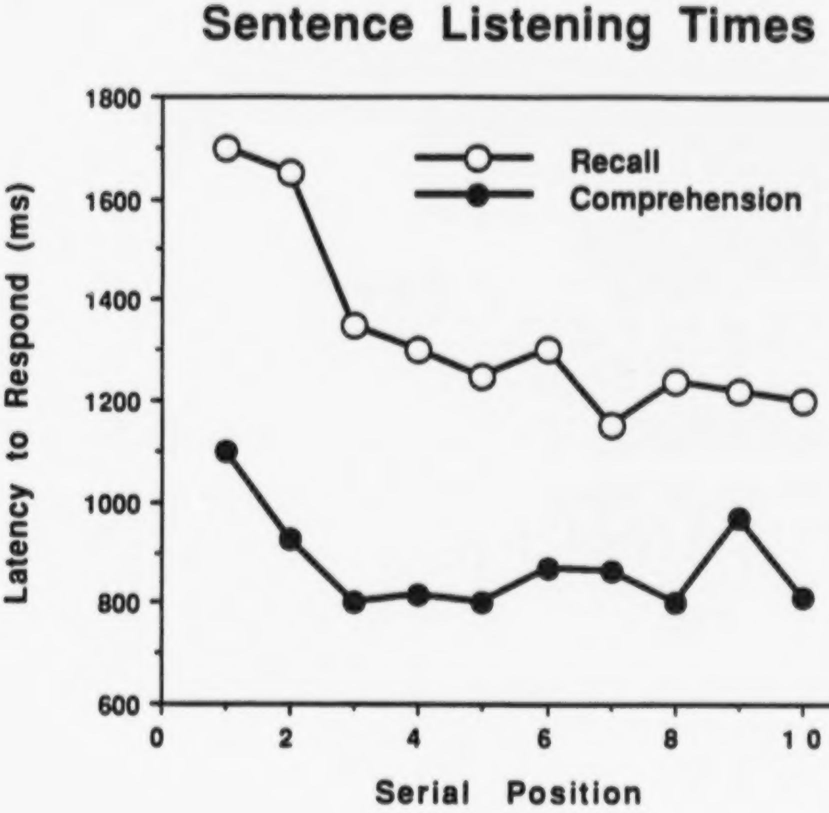


Figure 10. Sentence listening times for natural speech passages as a function of the serial position of sentences in the passage. Open circles represent latencies when subjects expected a subsequent verbatim recall test, and filled circles represent listening times when subjects expected a subsequent comprehension test (from Mimmack, 1982).

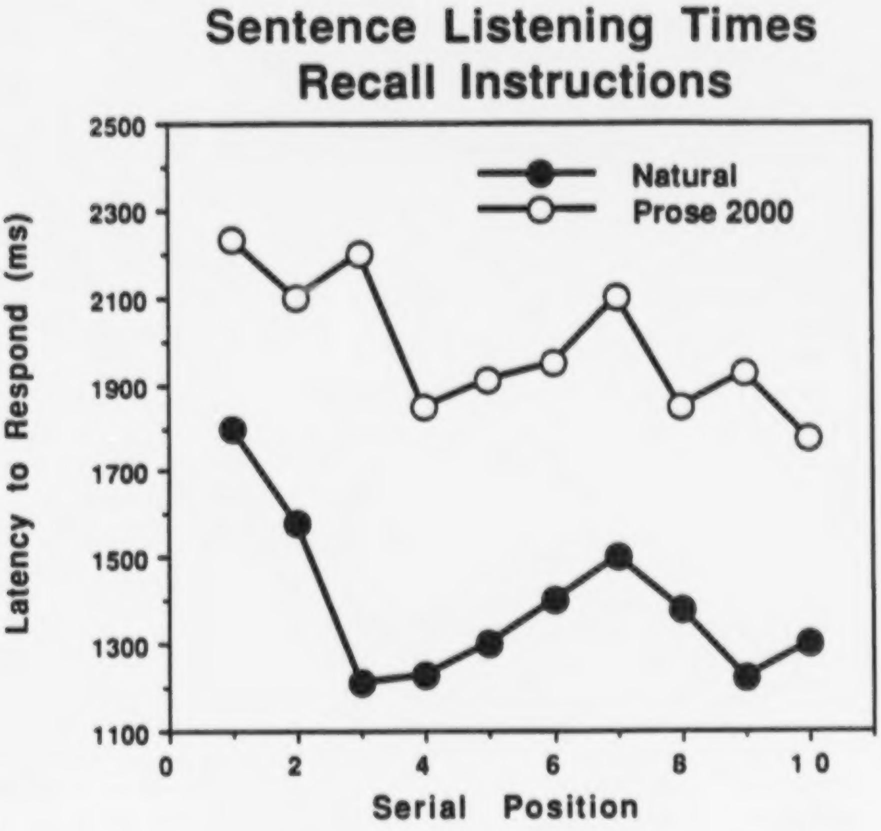


Figure 11. Sentence listenting times for natural and synthetic speech (Prose 2000) passages as a function of the serial position of sentences in the passage. Open circles represent listening times for the synthetic passages and closed circles represent listening times for natural passages. Subjects were expecting a verbatim recall memory task (from Mimmack & Pisoni, 1982).

speech (Ralston, Mullennix, Lively, Greene & Pisoni, 1989). The experiment was carried out to measure on-line differences in processing load during comprehension. Several short passages were produced either by a Votrax synthesizer or a natural talker. All passages were adapted from published reading comprehension tests for fourth grade and college level readers. Before a passage was presented, either 0, 2, 4, or 8 target words were displayed on a video monitor. Subjects were instructed to study and memorize the words for 30 seconds. When the study phase was completed, subjects were required to press a response button when they detected any of the target words in the passage. After each passage, subjects judged whether test sentences were true or false based on information contained in the preceding passage. Half of the sentences assessed memory for specific words occurring in the passage, and half probed memory for propositional information in the text base. The results showed that for every dependent measure, performance was always better for listeners who heard natural speech compared to those who heard synthetic speech.

Insert Figure 12 about here

Monitoring accuracy data are presented in Figure 12. Monitoring accuracy was higher for targets in natural passages than in Votrax passages. Accuracy decreased with increasing target set size. Thus, word monitoring performance in this task was affected by both signal quality and concurrent memory load. There was also a significant interaction between target set size and text difficulty. Accuracy was greater for the fourth grade passages only when subjects monitored for 8 word targets. This result is consistent with the hypothesis that retention of target items and the dynamics of comprehension both place processing demands on STM capacity.

Insert Figure 13 about here

Figure 13 displays monitoring latency data for correct detections as a function of voice and text difficulty. Monitoring responses were faster for the natural passages, demonstrating again that signal quality does affect the rate of comprehension. Although there was a main effect of target set size, post-hoc paired comparisons failed to reveal significant differences between the conditions. Monitoring responses were faster for the fourth grade passages than the college level passages, confirming our assumption that this linguistic variable loads STM. Finally, there was a significant interaction between voice and text difficulty. The increase in latency from fourth grade to college level text was larger for the Votrax passages than the natural passages. This result suggests that perceptual encoding and comprehension processes compete for common STM resources.

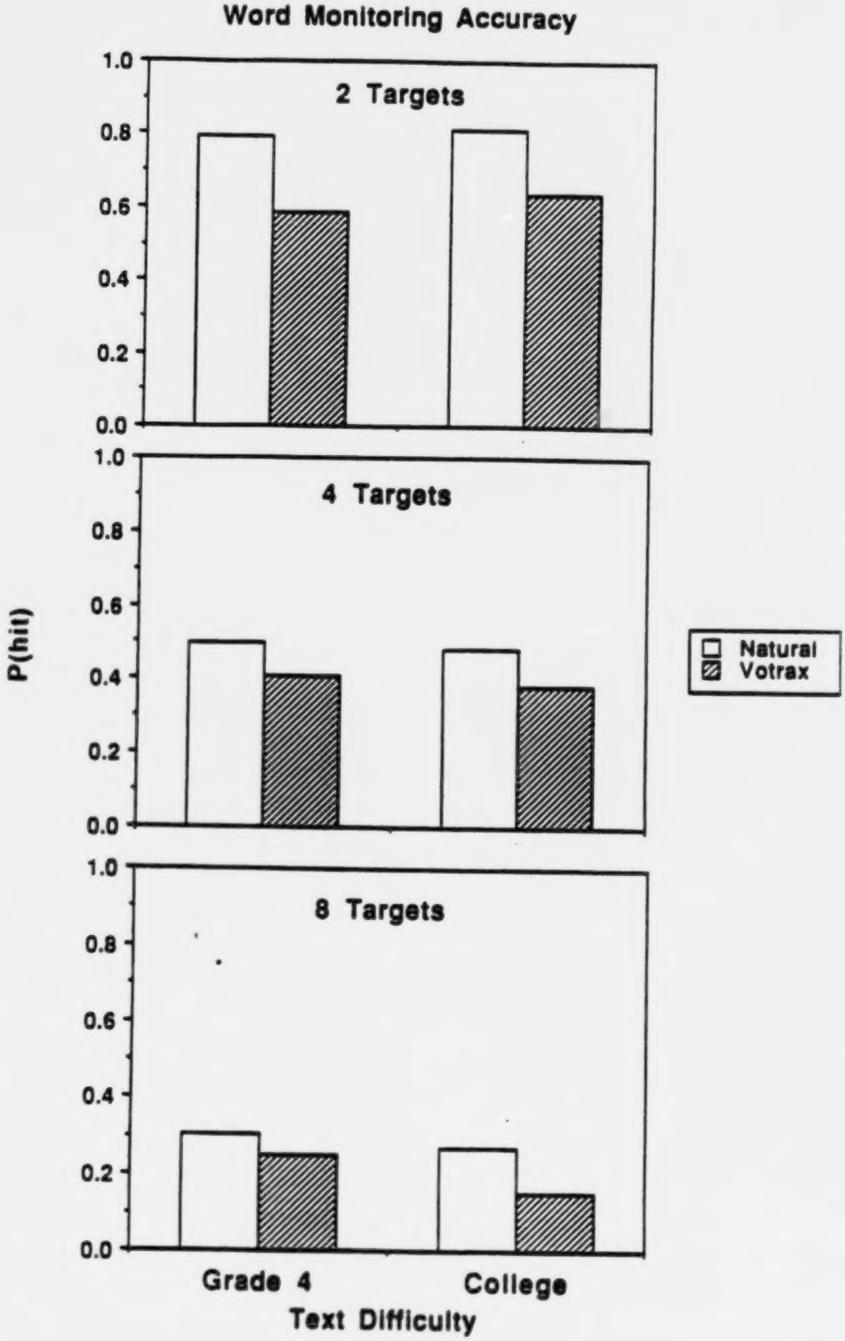


Figure 12. Word monitoring accuracy (probability of a hit) for two word target (upper panel), four word target (middle panel) and eight word target (bottom panel) conditions as a function of text difficulty. Open bars represent accuracy for natural speech passages and striped bars represent accuracy for synthetic speech (Votrax) passages (from Ralston et al., 1989).

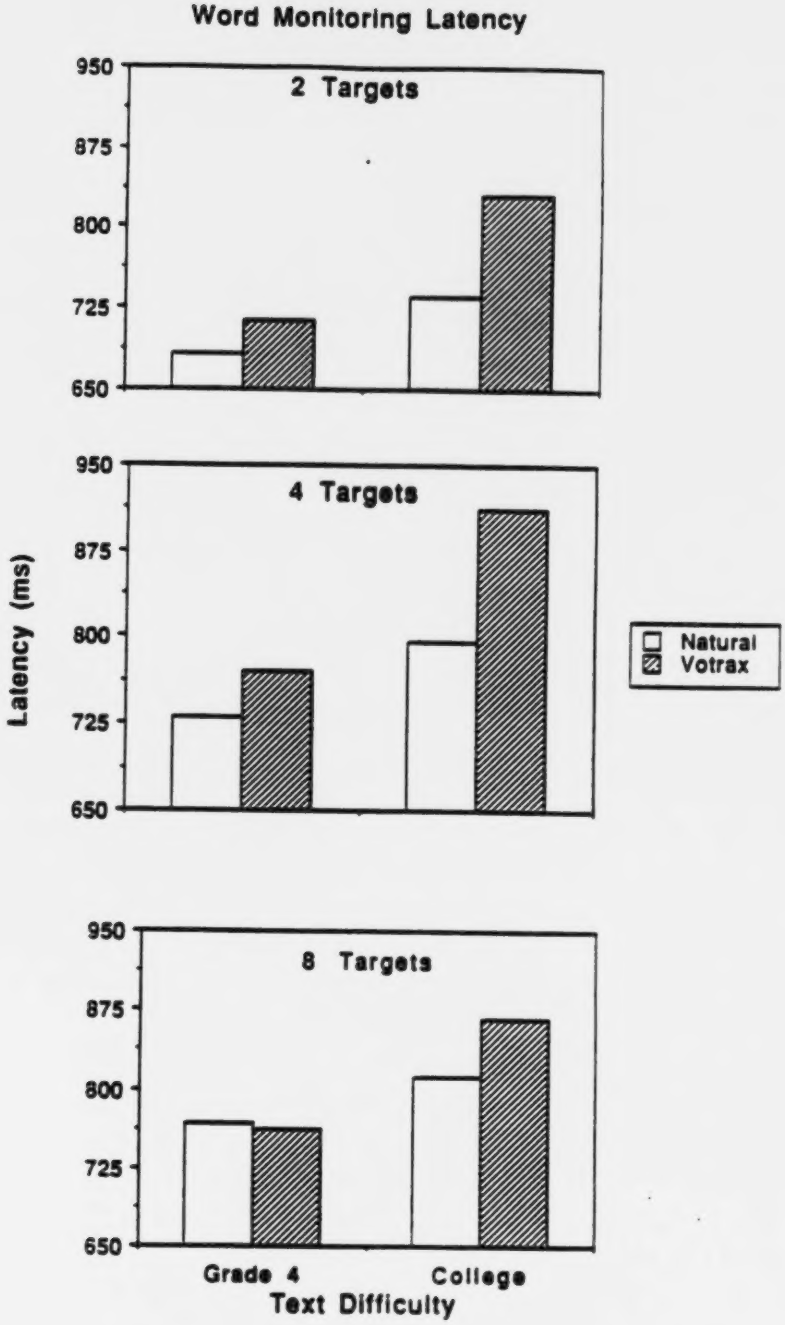


Figure 13. Word monitoring latency (in ms) for two word target (upper panel), four word target (middle panel) and eight word target (bottom panel) conditions as a function of text difficulty. Open bars represent latencies for natural speech passages and the striped bars represent latencies for synthetic speech (Votrax) passages (from Ralston et al., 1989).

Insert Figure 14 about here

Figure 14 displays sentence verification accuracy data as a function of voice and sentence type. Subjects verified sentences more accurately when they followed natural passages compared to synthetic passages. There was also an interesting interaction between voice and sentence type. While subjects listening to natural passages performed equally well on the two types of sentences, subjects listening to Votrax passages were less accurate on proposition recognition sentences. This result is similar to the earlier findings reported by Luce (1981).

Analysis of other factors indicated that accuracy was lower for the college level passages than the fourth grade passages. Although target set size was a significant factor in an ANOVA, subsequent analyses failed to reveal significant differences between the different monitoring conditions.

Insert Figure 15 about here

Figure 15 displays response latency data for correct verification responses as a function of voice and text difficulty. The largest influence on verification latency was clearly the type of sentence. Responses were nearly one second faster for word recognition sentences compared to proposition recognition. This result is consistent with the assumption that proposition recognition involves comprehension processes subsequent to lexical access.

In general, verification responses were faster for sentences following natural passages than synthetic passages. However, statistical analyses revealed that the voice factor was only marginally significant. Finally, there was a significant interaction between voice and text difficulty in the proposition data, displayed in the lower panel (voice x sentence type x text difficulty $[F(1,109) = 6.16, p = .01]$). Although response latencies were faster for sentences following natural passages, there was no latency difference between voices for the college level passages.

In summary, Ralston et al. found that the comprehension of poor quality synthetic speech is slower and less accurate than the comprehension of natural speech passages. The sentence verification data indicated that memory for linguistic information in passages of synthetic speech was also degraded in some way. The verification accuracy data also reveals that propositional information derived from passages of synthetic speech was particularly poor. This result provides support for an earlier speculation of Pisoni (1982) who suggested that subjects may listen and process synthetic speech differently than natural speech. In

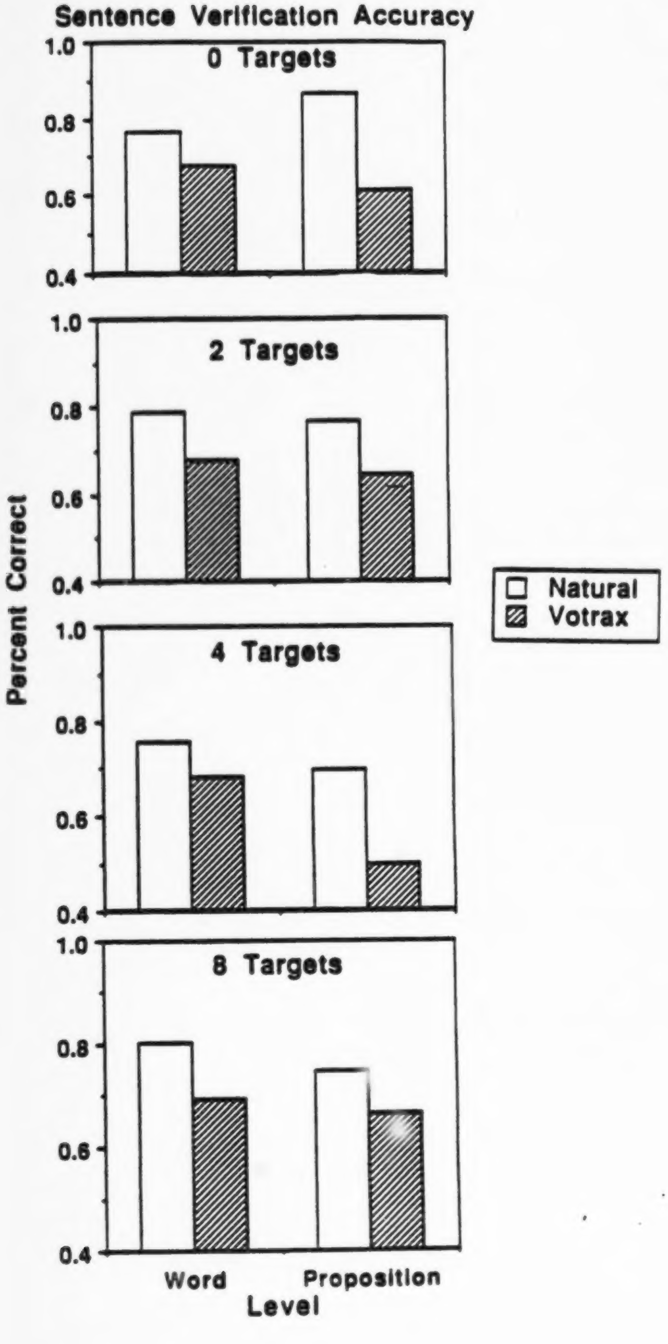


Figure 14. Sentence verification accuracy (probability correct) for zero, two, four and eight word target conditions (top to bottom panels, respectively) as a function of sentence type. Open bars represent accuracy for sentences following natural speech passages and striped bars represent accuracy for sentences following synthetic speech (Votrax) passages (from Ralston et al., 1989).

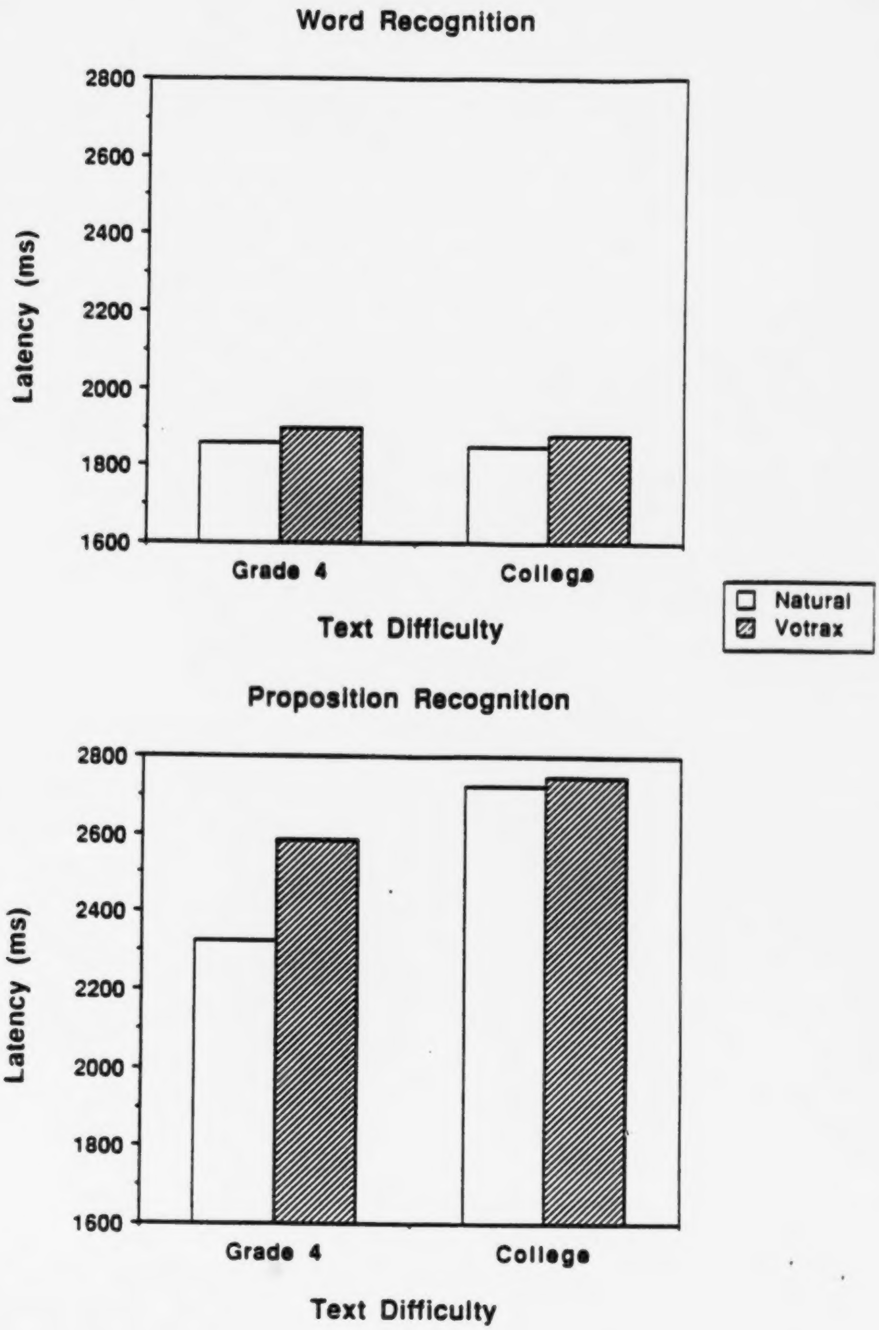


Figure 15. Sentence verification latency (in ms) for word recognition sentences (top panel) and proposition recognition sentences (bottom panel) as a function of text difficulty. Open bars represent verification latencies for sentences following natural speech passages, and striped bars represent verification latency for sentences following synthetic speech (Votrax) passages (from Ralston et al., 1989).

particular, listeners may allocate more attention to processing the acoustic-phonetic structure of synthetic speech than the content or meaning of the message. With more attention devoted to acoustic-phonetic processing, less resources are available to allocate to comprehension and memory processes, thus producing a less complete and less stable memory representation of the propositional information.

Summary and Conclusions

Although our knowledge about the comprehension of synthetic speech remains incomplete, several conclusions may be drawn from the studies reviewed above. Many other questions can be identified for future research. Based on the studies considered earlier, we examine the following issues: efficiency of comprehension, attentional demands, the role of perceptual learning and applications of the research. Finally, we discuss extensions of these findings and suggest several new areas for research.

Efficiency of Comprehension

Early comprehension studies failed to uncover reliable differences in performance between natural and synthetic speech (Jenkins & Franklin, 1981; McHugh, 1976; Nye et al., 1975; Pisoni & Hunnicutt, 1980; Schwab et al., 1985). However, more recent studies examining the comprehension of isolated sentences and connected discourse have reported significant differences in both accuracy and latency measures when more sensitive dependent measures are used (Luce, 1981; Moody & Joost, 1986; Manous et al., 1985; Pisoni et al., 1987). Several of these differences have been replicated and extended in the recent study carried out in our laboratory by Ralston et al. (1989) using an on-line word monitoring technique combined with a sentence verification task.

The failure of earlier studies to find significant differences in comprehension between natural and synthetic speech may be due to a variety of factors including selection of stimulus materials and specific testing techniques. For example, most studies employing passages of fluent connected speech relied on published multiple choice tests to assess comprehension. The accuracy of responding to these questions appears to be closely related to prior real-world knowledge. Such an effect would increase within-groups error variance and would serve to obscure any true differences in comprehension performance between conditions. This observation suggests the need for better experimental controls in comprehension research. One control procedure is to pre-screen test questions and discard those items that subjects can correctly answer without listening to the corresponding passages (Ralston et al., 1989). Although this technique works in a "statistical" sense over a number of subjects, there are no guarantees for all individuals. Because of this problem, other techniques may be preferable.

In particular, the use of passages referring to fictional events will increase the probability that subjects make use of information acquired through comprehension of test passages.

Results from sentence verification and word monitoring tasks indicate that synthetic speech is responded to more slowly than natural speech, even when the intelligibility of the two sets of materials is equated (Pisoni et al., 1987). The slower response times in comprehension may originate at an acoustic-phonetic or phonological level. The relatively impoverished cue structure of synthetic speech may induce slower phonological processing compared to natural speech. In addition, post-perceptual verification responses may also be slower after listening to synthetic speech. This overall pattern of results suggests that the memory representations of synthetic speech may be more fragile than natural speech. If this explanation is correct, memory for information derived from synthetic passages may decay more rapidly than information extracted from natural passages, even if the initial accuracy levels for the two sets of materials are equivalent. Studies of long-term retention of natural and synthetic speech need to be carried out to assess this hypothesis.

Capacity Demands

The results of earlier recall experiments suggested that perceptual encoding of synthetic speech incurs a greater processing load than natural speech (Luce et al., 1983). However, it is difficult to extrapolate these results directly to a comprehension task. To the extent that connected discourse is informationally redundant or otherwise easy to process, one would expect comprehension demands to be minimized. In this case, context-guided expectations might serve to neutralize or cancel the encoding demands of synthetic speech. Therefore, we would expect that the demands of synthetic speech would be more apparent as the complexity or difficulty of the text increase. Based on the available evidence, the effects of attentional load on comprehension are ambiguous, but suggest that the increased load does affect comprehension in some manner. Considering verification studies using isolated sentences, one report found an interaction between voice and load (Manous et al., 1985) while another did not (Pisoni et al., 1987). The two studies of passage comprehension involving the manipulation of cognitive load reported interactions with voice (Moody & Joost, 1986; Ralston et al., 1989). However, as noted above, there are reasons to view the Moody and Joost results cautiously. In contrast, Ralston et al. found that the increase in monitoring latencies observed with synthetic speech was greater for college-level passages, a result that is consistent with limited capacity expectations. Future research should examine this issue in greater detail because it has important ramifications for the application of speech I/O technology in high information/workload environments.

Training Effects

Another important issue deals with the effects of training on comprehension of synthetic speech. A number of studies have examined short-term training and have found reliable effects on comprehension (Jenkins & Franklin, 1981; McHugh, 1976; Pisoni & Hunnicutt, 1980). Several researchers have concluded that practice effects in comprehension reflect the learning of new mapping rules which relate the acoustic structure of a novel synthetic voice to known phonemic categories (Lee & Nusbaum, 1989; Schwab et al., 1985). This acoustic-phonetic relearning presumably involves processing costs that impact on other cognitive activities (Luce, 1981; Pisoni, 1982). Listeners may also learn perceptual segmentation strategies through training (Greenspan et al., 1988). However, Lee and Nusbaum (1989) recently suggested that processing synthetic speech does not require more attention than processing natural speech, but that listeners presented with synthetic speech initially mis-allocate resources. They argue that training only helps listeners efficiently re-allocate attention.

At the present time, no attempts have been made to train subjects to asymptote or to determine whether differences in comprehension between natural and synthetic speech will still be present after extensive training with these systems. However, studies utilizing connected discourse have implicitly assumed that training effects have been minimized by virtue of the explicit training at the beginning of testing sessions. Clearly, this is an empirical issue. In fact, no training studies, either with isolated words, sentences, or passages, have demonstrated that subjects reached a performance plateau. It is important to determine whether the attentional load imposed by encoding processes or by comprehension processes diminish with training or whether the load remains. Some cognitive processes may be automatized with training and place less demands on STM. Training studies have demonstrated that some individuals may learn to read passages for comprehension and take dictation at the same time (Hirst, Spelke, Reaves, Caharack & Neisser, 1980; Spelke, Hirst & Neisser, 1976). Similar trends may emerge in training studies with synthetic speech. Finally, if training is simply a matter of learning new acoustic-phonetic mappings, do subjects eventually reallocate attention to other processes once that learning is complete? If so, we would expect that the "levels of processing" effects observed with synthetic speech (Luce, 1981; Ralston et al., 1989) would be reduced after extensive amounts of training.

Data from the study using digitally-encoded speech by Schmidt-Nielson and Kallman (1987) suggested that subjects may make greater use of contextual information after a period of learning. The authors speculated that in these situations, subjects diverted their attention to sentential context, which in turn helped constrain acoustic-phonetic decisions. Future research should assess whether a similar process is operating when subjects listen to synthetic speech. If so, listeners may respond to synthetic speech with two successive strategies. When first exposed to synthetic speech, subjects might divert attention to the acoustic-phonetic structure of the signal. Once subjects acquire new acoustic-phonetic mapping rules, they learn how to locate word boundaries and identify words (Greenspan et al., 1988). As learning progresses, attention is diverted to processing contextual information which in turn constrains phonological decisions. This hypothesis can be tested in experiments designed to probe processing of different levels of information as a function of training and exposure to synthetic speech (see Lee & Nusbaum, 1989).

Applications

The results discussed in the present chapter suggest that the comprehension of connected synthetic speech is highly correlated with segmental intelligibility. This conclusion argues against the use of low- to moderate-quality synthetic speech for applications requiring very high levels of comprehension. The word monitoring and sentence verification data summarized above demonstrate that comprehension of synthetic speech proceeds at a slower rate than natural speech. Therefore, particularly low-quality synthetic speech may also be inappropriate for applications requiring rapid responses. Finally, several studies have shown that comprehension of connected synthetic speech may incur greater processing costs than natural speech. Thus, poor-quality synthetic speech also may be inappropriate for applications using difficult text, such as that of a data-base retrieval system, or with other competing tasks that place significant demands on working memory, such as in a cockpit, air traffic control tower, or battlefield management system.

All of these reservations may be modified after we gain a better understanding of the effects of training on comprehension. Many of the major effects associated with the comprehension of synthetic speech (lower accuracy, slower processing, and increased cognitive load) may be eliminated with appropriate practice (Lee & Nusbaum, 1989). If these effects disappear with training, the only limiting factors for applications would be practice and exposure.

Certain listener populations, such as young children or elderly adults, may experience difficulty comprehending synthetic speech (Greene & Pisoni, 1988). Capacity limitations and processing speed are known to be more constrained compared to college-age listeners who are typically used in these experiments (Greene & Pisoni, 1988; Salthouse, 1988). The same capacity arguments may apply to listeners or listening situations which degrade speech signals. This includes the hearing-impaired community, which is becoming increasingly dominated by older presbycutic listeners. Capacity limitations may also play an important role with signal degradations, such as noisy communication channels or reverberant environments.

Future Directions

Several problems still need to be studied. These may be grouped into the following major categories: capacity demands, training effects, memory decay and generalization.

Capacity Demands. As indicated above, the extent to which encoding demands compete with other comprehension processes for limited STM capacity is still a topic of great interest. While sentence verification experiments have found interactions between voice and memory load variables (such as sentence length) the evidence from studies conducted with fluent connected speech is equivocal. Therefore, further studies should be conducted with passage-length materials. These studies will help answer a number of basic questions (i.e., "Which cognitive mechanisms make use of limited attentional capacity?") as well as applications questions (ie., "Is the comprehension of synthetic speech compromised by difficult text or competing tasks?"). We consider here possible methodological improvements for investigating the role of text difficulty and competing cognitive tasks as they relate to capacity demands in comprehension. Either of these techniques may be combined with simultaneous measures to provide a more sensitive index of comprehension processes.

Previous studies utilizing passages of varying difficulty have also used different comprehension questions. One experimental strategy which circumvents this confound is to express the same propositional information in different texts varying in surface structure complexity. In this manner, the same post-passage sentences or questions could be used, yielding better control of textual difficulty.

The "digit preload" technique is a well-known method which may also be used to assess the extent to which the comprehension of synthetic speech competes with other cognitive tasks (Baddeley & Hitch, 1974; Lee & Nusbaum, 1989; Luce et al., 1983). For example, subjects could be required to memorize a variable-sized set of digits before listening to natural or synthetic passages for comprehension. Interactions between digit set size and voice would provide strong evidence that both variables draw on a common resource pool.

Another issue related to capacity limitations deals with listener fatigue or habituation. Assuming that there are increased processing demands for synthetic speech, one might expect that listeners would become mentally fatigued, and consequently that performance would decay more rapidly as a function of time. As a test, one could examine performance over time as a function of voice (see Mack, 1989, for preliminary data on this issue).

Training Effects. From an applications perspective, it is important to determine the time-course of training effects and asymptotic levels of comprehension performance as well as capacity demands. For example, a system that yields initially poor comprehension performance may, with exposure and training, ultimately yield high comprehension performance. This information, as well as costs associated with training, should be weighed with other factors when selecting a particular TTS system.

Memory Decay. The increased processing load for synthetic speech may also lead to poorer memory for comprehended information. This possibility is suggested by earlier studies demonstrating poorer STM retention of degraded natural (Dallett, 1964; Rabbitt, 1966) and synthetic word lists (Luce et al., 1983). In addition, there may be LTM losses for comprehended information even when original comprehension levels for synthetic speech are equivalent to those obtained for natural speech. Such a result would indicate a subtle, but important effect which should be considered in application decisions. A simple way to test this would be to administer comprehension tests at variable intervals after spoken passages to assess long-term retention. To our knowledge, studies of long-term retention have not been carried out yet with passages of synthetic speech.

Generalization. A more global issue that has received relatively little attention to date is the generality of the results obtained with synthetic speech. Are the results obtained with synthetic speech also found with other types of processed speech? At the level of segmental intelligibility, the answer is clearly "no." Nusbaum, Dedina & Pisoni (1984) demonstrated that perceptual confusions are dramatically different for DECtalk and noise-degraded natural speech CV syllables. We might expect that confusion patterns would vary with the various speech output devices, reflecting specific acoustic-phonetic synthesis rules for TTS systems, specific interactions between digital-encoding techniques and a speech signal, or interactions between different types of noise and a speech signal. However, other phenomena, such as capacity demands, listener fatigue, and "levels of processing effects" (Luce, 1981; Ralston et al., 1989; Schmidt-Nielson & Kallman, 1987) have not been investigated in studies with natural, synthetic, vocoded, and noise-degraded speech. Future experiments should determine whether these effects reflect general adaptive strategies of listeners to degraded or impoverished speech signals or whether they too are specific to the particular output device using synthetic speech.

Miscellaneous Issues. Evaluation studies of synthetic speech using special user populations have received greater attention recently (Greene & Pisoni, 1988; Mack, 1987, 1989; Ozawa & Logan, 1989). These have included studies with children and non-native English speakers. However, there have been no studies conducted with the elderly, a rapidly growing segment of the American population. Because of their increasing numbers, and because of the broader application of TTS devices, there is an increasing need to determine whether this population encounters special problems with synthetic speech. Due to the prevalence of presbycusis and attentional deficits in this population, we would expect that they would experience some problems in perception and comprehension.

Likewise, demanding or stressful environments, such as noisy or reverberant sound fields, should be studied. Because these environments may be relatively common in application settings, it is important to assess their impact on performance. Noisy and reverberant environments degrade the quality of speech signals, and therefore are expected to degrade perceptual performance. Part of the expected performance decrements may be due to relatively "low-level" effects, such as degrading the physical signal or auditory masking. However, there may also be more central contributions to performance decrements, particularly attentional constraints.

Finally, a greater emphasis should be placed on implementing voice technology using synthetic speech in real-world applications, especially when an operator is required to carry out several simultaneous cognitive tasks. Although laboratory studies have demonstrated differences in performance between natural and synthetic speech, there is a paucity of data collected in "natural" situations. Data-base query systems, computer-aided instruction, and voice mail are reasonable starting points for applied research on synthetic speech.

In summary, this chapter has reviewed research on the comprehension of synthetic speech carried out over the last 15 years. A large number of studies have demonstrated differences in segmental intelligibility between natural and synthetic speech. However, the evidence regarding comprehension is less conclusive. This is especially true with respect to comprehension of passages of fluent connected speech. Although the accumulated evidence to date indicates reliable differences in comprehension between natural speech and a wide variety of different kinds of synthetic speech, the results are quite variable across different studies, suggesting that important methodological factors need to be controlled before true differences in comprehension performance can be uncovered. In comprehension studies that have used successive measurement techniques, the results are equivocal. On the other hand, in comprehension studies that have used on-line simultaneous measurements, the results show reliable effects that are correlated with segmental intelligibility scores. When segmental intelligibility is equated between natural and synthetic speech, differences in comprehension performance have still been observed suggesting a general attenuation of the processes used to construct semantic and syntactic representations from acoustic-phonetic information contained in the speech signal. Additional research is needed to further understand the locus of these differences in comprehension performance. Because spoken language processing is extremely robust, it is often difficult to observe differences in comprehension without using fine-grained simultaneous measurement techniques.

References

Aaronson, D. & Scarborough, H. (1976). Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 42-55.

Abrams, K. & Bever, T.G. (1969). Syntactic structure modifies attention during speech perception and recognition. *Quarterly Journal of Experimental Psychology*, **21**, 280-290.

Allen, J., Klatt, D.H. & Hunnicutt, S. (1987). From *Text to speech: The MITalk system*. Cambridge, UK: Cambridge University Press.

Allen, J. (1981). Linguistic based algorithms offer practical text-to-speech systems. *Speech Technology*, **1**, 12-16.

Anderson, J.R. (1974). Verbatim and propositional representation of sentences in immediate and long-term memory, *Journal of Verbal Learning and Verbal Behavior*, **13**, 149-162.

American National Standards Institute (1969). *Methods for the calculation of the articulation index.*

Baddeley, A.D. & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation, Vol. 8.* New York: Academic Press.

Bartlett, F.C. (1932). *Remembering.* Cambridge: Cambridge University Press.

Blank, M.A., Pisoni, D.B. & McClaskey, C.L. (1981). Effects of target monitoring on understanding fluent speech. *Perception & Psychophysics*, **29**, 383-388.

Britton, B.K., Holdredge, T., Curry, C. & Westbrook, R.D. (1979). Use of cognitive capacity in reading identical texts with different amounts of discourse level meaning. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 262-270.

Brooks, L.R. (1968). Spatial and verbal components in the act of recall. *Canadian Journal of Psychology*, **22**, 349-368.

Brunner, H. & Pisoni, D. B. (1982). Some effects of perceptual load on spoken text comprehension. *Journal of Verbal Learning and Verbal Behavior*, **21**, 186-195.

Cairns, H.S. & Kamerman, J. (1975). Lexical information processing during sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **14**, 170-179.

Carr, T.H. (1986). Perceiving visual language. In K.R. Boff, L. Kaufman & J.P. Thomas (Eds.), *Handbook of perception and human performance, Vol. II.* New York: Wiley.

Chial, M.R. (1984). Comparison of commercial speech synthesizers for small computers. Paper presented at the 1984 ASHA annual convention, San Francisco, Nov. 1984.

Cirilo, R.K. & Foss, D. (1980). Text structure and reading time for sentences. *Journal of Verbal Learning and Verbal Behavior*, **19**, 96-109.

Clark, H.H. & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, **3**, 472-517.

Collins, A.M. & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **8**, 240-247.

Crowder, R.G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, **20**, 55-60.

Dallett, K.M. (1964). Intelligibility and short-term memory in the repetition of digit strings, *Journal of Speech and Hearing Research*, **7**, 362-368.

Egan, J.P. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test. *Journal of the Acoustical Society of America*, **30**, 596-600.

Fairbanks, G., Guttman, N. & Miron, M.S. (1957) The effects of time compression upon the comprehension of connected speech. *Journal of Speech and Hearing Disorders*, **22**, 10–19.

Foss, D.J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, **8**, 457-462.

Foss, D.J. (1970). Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **9**, 699-706.

Foss, D.J. & Lynch, R.H. (1970). Decision processes during sentence comprehension: Effects of surface structure on decision times. *Perception & Psychophysics*, **5**, 145-148.

Gough, P.B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, **4**, 107-111.

Gough, P.B. (1966). The verification of sentences: The effects of delay of evidence and sentence length. *Journal of Verbal Learning and Verbal Behavior*, **5**, 492-296.

Greene, B.G. & Pisoni, D.B. (1988). Perception of synthetic speech by adults and children: Research on processing voice output from text-to-speech systems. In L.E. Bernstein (Ed.), *The vocally impaired: Clinical practice and research*, Philadelphia: Grune & Stratton.

Greenspan, S.L., Nusbaum, H.C. & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 421-433.

Guralnik, D.B. (1986). *Webster's new world dictionary of the American language*, New York: Prentice.

Haggard, M., Ambler, S. & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, **47**, 613-617.

Hakes, D.T. & Foss, D.J. (1970). Decision processes during sentence comprehension: Effects of surface structure reconsidered. *Perception & Psychophysics*, **8**, 413-416.

Hersch, H.M. & Tartarglia, L. (1983). *Understanding synthetic speech*. User Research Group, Corporate Research & Architecture, Maynard, MA: Digital Equipment Corporation.

Hirst, W., Spelke, E.S., Reaves, C.C., Caharack, G. & Neisser, U. (1980). Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General*, **109**, 98-117.

Hoover, J., Reichle, J., van Tasell, D., & Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax, *Journal of Speech and Hearing Research*, **30**, 425-431.

House, A.S., Williams, C.E., Hecker, M.H. & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

Huggins, A.W. & Nickerson, R.S. (1985). Speech quality evaluation using "phoneme-specific" sentences. *Journal of the Acoustical Society of America*, **77**, 1896-1906.

Jenkins, J.J. & Franklin, L.D. (1981). Recall of passages of synthetic speech. Paper presented at the Psychonomics Society Meeting, November, 1981.

Johnston, P.H. (1984). Assessment in reading. In Pearson, P.D. (Ed.), *Handbook of reading research*, New York: Longman.

Kalikow, D., Stevens, K. & Elliot, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, **61**, 1337-1351.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Kintsch, W. & Keenan, J.M. (1973). Reading rate as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**, 257-274.

Kintsch, W. & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363-394.

Kintsch, W., Mandel, T.S. & Kozminsky, E. (1977). Summarizing scrambled stories. *Memory & Cognition*, **5**, 547-552.

Kintsch, W., Kozminsky, E., Sterby, W. J., McKoon, G. & Keenan, J.M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, **14**, 196-214.

Klapp, S.T. & Netick, A. (1988). Multiple resources for processing and storage in short-term working memory. *Human Factors*, **30**, 617-632.

Klatsky, R.L. (1980). *Human memory*, San Francisco: Freeman.

Klatt, D.H. (1979). Speech perception: A model of acoustic- phonetic analysis and lexical access. In R.A. Cole (Ed.), *Perception and production of fluent speech*, Hillsdale, NJ: Erlbaum.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737–793.

Klatt, D.H. (1988). Review of selected models of speech perception. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process*, Cambridge: MIT Press.

Lee, L. & Nusbaum, H.C. (1989). The effects of perceptual learning on capacity demands for recognizing synthetic speech. *Journal of the Acoustical Society of America*, **85**, YY10.

Levelt, W.J.M. (1978). A survey of studies in sentence perception. In W.J.M. Levelt & G.B. Flores d'Arcais (Eds.), *Studies in the perception of language*, New York: Wiley.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Logan, J.S., Greene, B.G. & Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566-581.

Logan, J.S. & Pisoni, D.B. (1986). Intelligibility of phoneme specific sentencess using three text-to-speech systems and a natural speech control. *Research on speech perception progress report no. 12*, (pp. 319-334). Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A. (1981). Comprehension of fluent synthetic speech produced by rule. *Research on speech perception progress report no. 7*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A., Feustel, T.C. & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17-32.

McClelland, J.L. & Elman, J.L. (1986). Interactive processes in speech perception. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing*, Volume 2, Cambridge: MIT Press.

McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Naval Research Laboratoroy Report 8015.

Mack, M. (1987). Perception of natural and vocoded sentences among English monolinguals and German-English bilinguals. *Journal of the Acoustical Society of America*, Suppl.1, **81**, A16.

Mack, M. (1989). The intelligibility of LPC-vocoded words and sentences presented to native and non-native speakers of English. *Journal of the Acoustical Society of America*, Suppl.1, **86**, NN8.

Marslen-Wilson, W. & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1- 71.

Manous, L.M., Pisoni, D.B., Dedina, M.J. & Nusbaum, H.C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. *Research on speech perception progress report no. 11*, (pp. 33-58). Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Miller, G.A., Heise G.A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329-335.

Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, **2**, 217-228.

Mimmack, P.C. (1982). Sentence-by-sentence listening times for spoken passages: Text structure and listener's goals. Unpublished M.A. Thesis, Indiana University.

Mimmack, P.C. & Pisoni, D.B. (1982). Unpublished data.

Moody, T.S. & Joost, M.G. (1986). Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. *Proceedings of the Voice Input/Output Society*, Alexandria, VA, 1986.

Morton, J. & Long, J. (1976). Effects of word transition probability in phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **15**, 43-51.

Murdock, B.B. Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, **64**, 482-488.

Navon, D. (1984). Resources - A theoretical soupstone? *Psychological Review*, **91**, 216-234.

Navon, D. & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, **86**, 214-255.

Newell, A. & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norman, D.A. & Bobrow, D.G. (1975). On data-limited and process- limited processes. *Cognitive Psychology*, **7**, 44-64.

Nusbaum, H.C, Dedina, M.J. & Pisoni, D.B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. *Research on speech perception progress report no. 10*, (pp. 409-422). Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Nye, P.W. & Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). *Haskins Laboratory Status Report on Speech Research SR-33*, pp. 77-91. New Haven, CT: Haskins Laboratories.

Nye, P.W. & Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories: Status Report on Speech Research, SR-38.* New Haven, CT: Haskins Laboratories.

Nye, P.W., Ingemann, F. & Donald, L. (1975). Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. *Haskins Laboratories: Status Report on Speech Perception SR-41.* New Haven, CT: Haskins Laboratories.

Ozawa, K. & Logan, J.S. (1989). Perceptual evaluation of two speech coding methods by native and non-native speakers of English. *Computer Speech and Language*, **3**, 53-59.

Pisoni, D.B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology*, **1**, 10-23.

Pisoni, D.B. & Dedina, M.J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report. *Research on speech perception progress report no. 12*, pp. 3-18. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Pisoni, D.B. & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. *IEEE International Conference Rec. on Acoustics, Speech, and Signal Processing*, (pp. 572-575).

Pisoni, D.B. & Luce, P.A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, **25**, 21-52.

Pisoni, D.B., Manous, L.M. & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on sentence verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

Pisoni, D.B., Nusbaum, H.C. & Greene, B.G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **11**, 1665-1676.

Pisoni D.P. & Sawusch J.R., (1975). Some stages of processing in speech perception. In A. Cohen & S.G. Nooteboom (Eds.), *Structure and process in speech percpetion.* Berlin: Springer-Verlag.

Pollack, I. & Decker, L.R. (1958). Confidence ratings, message reception, and the receiver operating characteristic. *Journal of the Acoustical Society of America*, **30**, 286-292.

Pratt, R.L. (1987). Quantifying the performance of text-to-speech synthesizers. *Speech Technology*, **5**, 54-63.

Rabbitt, P. (1966). Recognition: Memory for words correctly heard in noise. *Psychonomic Science*, **6**, 383-384.

Ralston, J.V., Mullennix, J.W., Lively, S.E., Greene, B.G. & Pisoni, D.B. (1989). Comprehension of natural and synthetic speech. *Journal of the Acoustical Society of America*, Suppl.1, **86**, NN9.

Sachs, J.S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, **2**, 437-442.

Salthouse, T.A. (1988). The role of processing resources in cognitive aging. In M.L. Howe & C.J. Brainerd (Eds.), *Cognitive development in adulthood.* New York: Springer-Verlag.

Savin, H.B. & Bever, T.G. (1970). The non-perceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, **9**, 295-302.

Schmidt-Neilson, A. (this volume). Intelligibility testing for speech technology. In Bennett, Syrdal & Greenspan (Eds.), *Behavioral aspects of speech technology: Theory and applications.* New York: Elsevier.

Schmidt-Neilson, A. & Kallman, H.J. (1987). Evaluating the performance of the LPC 2.4 kbps processor with bit errors using a sentence verification task. NRL Report No. 9089, Washington, D.C., Naval Research Laboratory.

Schwab, E.C., Nusbaum, H.C. & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, **27**, 395-408.

Shields, J.L., McHugh, A. & Martin, J.G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, **102**, 250-255.

Shiffrin, R.M. (1976). Capacity limitations in information processing, attention, and memory. In W.K. Estes (Ed.), *Handbook of learning and cognitive process*. Hillsdale, N.J.: Erlbaum.

Shiffrin, R.M. (1987). Attention. In R.C. Atkinson, R.J. Hernstein, G. Lindsey, & R.D. Luce (Eds.), *Steven's handbook of experimental psychology*. New York: Wiley.

Schneider, W. & Shiffrin, R.M. (1977). Controlled and automatic humans information processing: I. Detection, search, and attention. *Psychological Review*, **84**, 1-66.

Spelke, E.S., Hirst, W.C. & Neisser, U. (1976). Skills of divided attention. *Cognition*, **4**, 215-230.

Streeter L.A. & Bever, T.G. (1975). The effects on the detection of linguistic and nonlinguistic stimuli are opposite at the beginning and end of a clause. Mimeographed, Columbia University. Cited in Leveldt, W.J.M. and Flores, G.B. (Eds.), *Studies in the perception of language*. New York: Wiley and Sons.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, **30**, 415-433.

Voiers, W. D. (1983). Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology*, **1**, 30-39.

Waugh, N.C. & Norman, D.A. (1965). Working memory. *Psychological Review*, **72**, 89-104.

Wickens, C.D. (1987). Information processing, decision making, and cognition. In Salvendy, G. (Ed.), New York: Wiley Interscience.

Wingfield, A. & Klein, J.F. (1971). Syntactic structure and acoustic pattern in speech perception. *Perception & Psychophysics*, **9**, 23-25.

Effects of Talker Variability on Speech Perception by 2-month-old Infants

Peter W. Jusczyk[1], David B. Pisoni and John Mullennix[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] Department of Psychology, University of Oregon and Laboratoire de Sciences Cognitives et Psycholinguistique, C.N.R.S., Paris.

[2] Department of Psychology, Wayne State University, Detroit, MI.

Abstract

The present study explores the way that talker variation influences the 2-month-old infants' perception and memory for speech sounds using the HAS procedure. Experiment 1 focuses on the consequences that talker variation has on the infant's ability to detect differences between speech sounds. When listening to versions of a syllable, such as /bʌg/, produced by six male and six female talkers, infants were able to detect a change to another syllable, such as /dʌg/, uttered by the same group of talkers. In fact, as far as detecting the difference was concerned, infants exposed to multiple talkers proved to be as good as other infants , who heard utterances produced by only a single talker. Moreover, results from other test conditions showed that infants could discriminate between the voices of the individual talkers, although discriminating one mixed group of talkers (three males and three females) from another proved to be too difficult for them. Experiment 2 explored the consequences of talker variation on infants' memory for speech sounds. The HAS procedure was modified by introducing a 2-minute delay period between the preshift and postshift phases of the experiment. In this condition, talker variation impeded the encoding of speech sounds by infants. However, infants who heard versions of the same syllable produced by 12 different talkers did not detect a change to a new syllable produced by the same talkers after the delay period. Infants who heard the same syllable produced by a single talker were able to detect the phonetic change after the delay. Finally, although infants who heard productions from a single talker retained information about the phonetic structure of the syllable during the delay, they apparently did not retain information about the identity of the talker. Experiment 3 demonstrated that talker variation need not interfere with the retention of all speech information by infants. Specifically, infants were able to recognize a change in the gender of the talkers' voices (from male to female or vice versa) after a 2-minute delay, even when six different males and six different females produced the sounds. These results have important implications for the way that word recognition processes and the mental lexicon may develop during language acquisition. Parallels are also noted in the way that talker variation affects speech processing by infants and adults.

Effects of Talker Variability on Speech Perception by 2-month-old Infants

One important aspect of language acquisition that begins to unfold during the first year of life is the development of a lexicon in the native language. Just as children begin to produce their first words towards the end of the first year, so too do they begin to understand words from their native language (e.g., Huttenlocher, 1974). Comprehending words requires that the infant store away some representation of the sound structure of the word so that they can retrieve the appropriate meaning. A number of the prerequisites necessary for the successful storage of words, and hence, for the development of the lexicon have been studied during the last 20 years of research on infant speech perception. For example, the capacities of infants to discriminate subtle phonetic distinctions have been well-documented (e.g., Aslin, 1987; Aslin, Pisoni & Jusczyk, 1983; Eimas, 1982; Jusczyk, 1981; Kuhl, 1987). In addition, a number of studies have shown that, by 6 months, infants are apparently able to ignore the variability in the speech signal introduced when the same item is uttered by different talkers. This latter ability is critical for being able to recognize the same word spoken by different individuals. Some recent work also demonstrates that even newborn infants apparently have some minimal capacity to represent different speech sounds (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy & Mehler, 1988; Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy & Mehler, in press). Nevertheless, many other important factors related to lexical development have yet to be explored. For example, very little is known about the capacity of infants to retain information about the speech sounds that they hear, or, indeed, about factors that might affect the retention of information about the sound properties of words. Information about such issues is critical in order to understand the way that the lexicon is structured and how it develops.

One factor known to affect the way that adults encode speech sounds is talker variability. Although adults are readily able to adjust for differences in talker's voices in perceiving speech sounds (e.g., Bladon, Henton & Pickering, 1984; Dechovitz, 1977; Disner, 1980; Fourcin, 1968; Gerstman, 1968; Neary, 1978; Rand, 1971; Summerfield 1975; Syrdal & Gopal, 1986; Verbrugge, Strange, Shankweiler & Edman, 1976), these kinds of adjustments are not without consequence for perceptual processing. Thus, the accuracy with which items are identified suffers when talker voice varies as opposed to when it remains constant (Creelman, 1957; Fourcin, 1968; Verbrugge et al., 1976). Similarly, the latencies required to perform identification (e.g., Summerfield, 1975; Summerfield & Haggard, 1973) and matching tasks (Allard & Henderson, 1975; Cole, Coltheart & Allard, 1974) have been shown to increase significantly when listeners are required to adjust to different talkers' voices. With respect to perception, the consequences associated with adjusting to different talkers' voices appear to be confined to early stages of acoustic-phonetic processing as opposed to higher level ones (Mullennix, Pisoni & Martin, 1989). Thus, Mullennix et al. showed that talker variability interacts with variables that affect acoustic-phonetic encoding (such as the presence of white noise) but not with variables that affect higher-level word recognition processes (such as lexical density and word frequency). However, there is also evidence that talker variation has consequences for cognitive processes other than those involved in perception. On the one hand, Martin, Mullennix, Pisoni & Summers (1989) found that memory processes in **recall tasks** can be adversely affected when listeners must cope with talker variation. In particular, their results suggested that both encoding processes and the efficiency of rehearsal processes used to transfer items into long term memory were disrupted by talker variation. On the other hand, Craik and Kirsner (1974) found that talker variability could actually have a beneficial effect on **recognition memory** of items presented in a list. Thus, their subjects were faster and more accurate for items that were repeated in the same voice as the original item.

In summary, despite the fact that adult listeners are able to adjust fairly rapidly to talker variation, there are indications that costs are associated with the process. These costs show up both with respect to the initial perceptual processing of the signal and in its encoding into long-term memory. It is also clear that information about talker differences are detected in perception and may be retained in memory. Although costs may be incurred in terms of the amount of information that can be encoded when talker variation occurs, there are some benefits as well for subsequent recognition of the items that are encoded.

As noted earlier, there is evidence that, by 6 months of age, infants display some basic ability to cope with talker variation. Kuhl (1979; 1983) showed that 6-month-olds will continue to detect a phonetic contrast in the face of changes in speaking voices that range from children to adults and include both males and females. Thus, infants trained to distinguish a contrast between two vowel tokens produced by a single talker successfully generalized this distinction to vowel tokens produced by different talkers, even when there was considerable acoustic overlap among the tokens of the distinctive vowel classes (Kuhl, 1983). Of course, were it the case that infants succeed because they are simply unable to distinguish differences between talkers' voices, then their achievement on this type of task would not be very remarkable. However, there is ample evidence to believe that this is not the case. Studies with newborn infants show them to be capable of recognizing their mothers' voices from those of other mothers (e.g., DeCasper & Fifer, 1980; Mehler, Bertoncini, Barriere & Jassik-Gerschenfeld, 1978; Mills & Meluish, 1974). Moreover, 6-month-olds are able to perform a task that requires responding to tokens produced by a particular talker as opposed to another talker (Miller, Younger & Morse, 1982).

Nevertheless, the question as to whether the infant's success at coping with talker variation also bears a cost for perceptual processing and memory has not been directly addressed in previous research. In fact, aside from the two studies by Kuhl (1979; 1983), the only attempt to focus on the way that infants handle irrelevant variation in speech sounds was a study by Kuhl and Miller (1982) with infants 1- to 4-months of age. Kuhl and Miller used synthetic vowel stimuli and examined the capacity of infants to detect a change in one dimension when a second dimension varied irrelevantly. The two dimensions were pitch and vowel quality. Their results indicated that when pitch varied irrelevantly, the infants were able to detect a vowel change. However, the converse did not hold. Namely, when infants were exposed to a series of randomly alternating vowels, /a/ and /i/, they did not detect a subsequent change in pitch contour. Consequently, Kuhl and Miller interpreted this as an indication that the vowel quality dimension was more salient for the infants and distracted them from detecting the change in pitch quality. As further support of their interpretation, Kuhl and Miller noted that infants took significantly longer to habituate to the stimuli when vowel quality varied irrelevantly than they did when pitch quality varied, suggesting that infants attended more to the vowel variation than to the pitch variation (however see Carrell, Smith and Pisoni, 1981).

Kuhl and Miller's findings are an interesting demonstration that irrelevant variation along some dimension may hinder infants from detecting a change along another less salient dimension. Still, it is hard to predict whether the type of variation introduced by alternating different vowels is of the same order of magnitude as one stemming from the presence of different talkers (Carrell et al., 1981). In fact, the studies with older infants indicate that coping with talker variation does not prevent infants from discriminating a contrast between different vowels (Kuhl, 1979; 1983). Nevertheless, what is not known at present is the extent to which infants may incur the kinds of subtle costs in processing and encoding speech that have been reported for adult listeners when dealing with talker variation (e.g. Martin et al., 1989; Mullennix et al., 1989). Information about the way that talker variation influences speech processing by infants is important not only for determining how the lexicon develops, but also for understanding the mechanisms that underlie the process of perceptual normalization. For instance, it has been suggested that normalization may operate at early stages of speech processing in a mandatory fashion, independently of higher-level cognitive processes (Miller, 1987; Mullennix et al., 1989). If so, then in line with other features that are associated with modular systems, one might expect to find that the characteristics of the normalization system are innately wired and fixed. Hence, strong parallels would be predicted for the way that talker variation influences perceptual processing in infants and adults.

It was with these issues in mind that the present study was undertaken. Accordingly, we designed a series of experiments to evaluate the way that talker variation affects the processing and retention of speech sounds by 2-month-old infants. The first experiment focuses on potential effects involved with the perception of a speech contrast. The second and third experiments are concerned with the impact that talker variation has on infants memory for speech information.

Experiment 1

Previous investigations of infants' capacity for dealing with talker variation have focused on infants 6-months of age (e.g., Kuhl, 1979; 1983). Hence, nothing is currently known about the capacity of younger infants to cope with talker variability. Consequently, we decided to use a modified version of the high-amplitude sucking (HAS) procedure to explore this capacity in 2-month-olds. The speech contrast that was selected involved a change in the initial consonant of two CVC syllables, /bʌg/ and /dʌg/ (corresponding to the English words, "bug" and "dug"). The syllable tokens for the study were chosen from six male and six female talkers who produced them originally for use in the Mullennix et al. (1989) study.

Determining the consequences of talker variation requires comparisons between situations involving single-talkers and comparisons involving multiple-talkers. For this reason, we decided to examine the same contrast in both the single- and multiple-talker conditions. Hence, one experimental group and one control group was tested with tokens from a single-talker. The experimental group was habituated to one of the syllables, either /bʌg/ or /dʌg/, and were presented with the remaining syllable during the postshift phase. The control group was habituated to one of the two syllables and continued to hear the same one during the test phase. There were two comparable multiple talker conditions. The only difference was that tokens from all 12 talkers were used during both the habituation and test phases of the experiment. By comparing the performance of the infants in the multiple-talker conditions with that of the infants in the single-talker conditions, we could evaluate the consequences of talker variation on the 2-month-old's capacity to detect a phonetic change. We hypothesized that any increase in processing load associated with the multiple-talker condition might show itself either in discrimination performance or in the time that it took for infants to habituate to the syllable(s) during the first phase of the procedure, as Kuhl and Miller reported for their study.

Two additional conditions were also included in the study. First, to determine whether the tokens from the different talkers used in the present study were discriminable for the infants, we tested a group in which the contrast was not a phonetic change, but rather a difference between two talkers. A number of previous studies have examined the ability of infants to detect differences between talkers' voices, but these studies used speech samples longer than a single syllable (e.g., DeCasper & Fifer, 1980; Kaplan, 1969; Turnure, 1971). In the present investigation, the syllable type (e.g., /dʌg/) was the same for both phases of the experiment, but the identity of the talker was changed after habituation to the first syllable. The remaining test condition was one that involved habituating the infants with tokens of a particular syllable type (e.g., /bʌg/) spoken by a set of 6 different talkers (3 males and 3 females). Then, following habituation to these tokens, the infants were switched to an entirely new set of talkers (3 males and 3 females) uttering the same syllable. The purpose of this last condition was to assess possible limitations on infants' abilities to encode information about talker identity. Thus, to discriminate the contrast in this last condition, infants would have to encode the syllables according to the identity of the talker and retain this information for comparison with the new tokens presented after habituation. Previous work with infants at this age indicates that they are capable of representing information about the phonetic content of syllables (e.g., /bi/, /ba/, bʌ/) so as to detect the presence of new syllable types (e.g., /bu/) presented after habituation (e.g., Bertoncini et al., 1988;

Jusczyk et al., in press; Jusczyk & Derrah, 1987). However, little is known about whether they also include information about talker identity in their representations of these syllables.

Method

Procedure. Each infant was tested individually in a small laboratory room. The infant was placed in a reclining chair facing a blank wall approximately 1 m. away. An image of flowers was projected on the wall for the entire test session. The picture was situated just above a loudspeaker through which the test stimuli were played. Each infant sucked on a blind nipple held in place by an experimenter who wore headphones and listened to recorded music throughout the test session. A second experimenter in an adjacent room monitored the test apparatus.

The experimental procedure was a modification of the high-amplitude sucking technique (Eimas, Siqueland, Jusczyk & Vigorito, 1971; Jusczyk, 1985b; Siqueland & DeLucia, 1969). For each infant, the high amplitude sucking criterion and the baseline rate of high amplitude sucking were established prior to the presentation of any test stimuli. The criterion for high-amplitude sucking was adjusted to produce rates of 15-35 sucks/min. After a baseline rate was established, the presentation of stimuli was made contingent on the rate of high-amplitude sucking. Criterion sucks resulted in the presentation of one speech syllable. For infants in the single-talker conditions, the same syllable was presented throughout the preshift phase of the experiment. For infants in the multiple-talker conditions, the syllables were selected at random from a set stored on a computer disk. Thus, it was possible that an infant in a multiple-talker condition might hear the same syllable or a different one for successive criterion sucks. The maximum stimulus presentation rate was one syllable per second. If the infant produced a burst of sucking with interresponse times less than 1 second, then each response did not produce one presentation of a stimulus. Instead the timing was reset so as to provide continuous auditory feedback for one second after the last response of the sucking burst. In any case, if the 1 sec. period would have terminated in the middle of a syllable, it was delayed until the syllable was completed.

The criterion for habituation during the preshift phase of the experiment was a decrement in sucking rate of 25% or more over 2 consecutive minutes compared with the rate in the immediately preceding minute. At this point, the auditory stimulation was changed to match that used in the postshift phase of a given condition. For infants in the experimental conditions, this resulted in a change in the stimuli presented. Infants in the control conditions continued to hear the same stimuli as before. The postshift phase began with the presentation of the first stimulus after the habituation criterion had been achieved. The infants' sensitivity to changes in auditory stimulation was inferred from comparisons of response rates of subjects in the experimental and control conditions during the postshift period. The postshift period lasted for at least 4 minutes or until the infant showed a 25% decrease in sucking for two consecutive minutes.

Stimuli. The stimuli consisted of natural versions of the syllables /bʌg/ and /dʌg/ produced by six male and six female talkers from Indiana. The stimuli were words recorded on audio tape in a sound attenuated booth using an Electro-Voice Model D054 microphone and a Crown 800 series tape recorder. The utterances were subsequently digitized via a 12-bit analog-to-digital converter and stored on a PDP 11/34 computer at the Speech Research Laboratory at Indiana University. The digitized versions of these stimuli were copied on floppy disk and transferred to a PDP 11/73 computer at the Speech Perception Laboratory at the University of Oregon. The stimuli were converted to analog form in real-time via a 12-bit digital-to-analog converter. They were accessed directly during the course of the experiment and played out through a 4.8 kHz low-pass filter. All of the words used in the experiment had been previously tested for intelligibility using a group of adult listeners at Indiana University. The items received identification scores of 95% correct or above when presented in isolation.

Design. Each infant was seen for one experimental session. Twelve subjects were randomly assigned to each of six test conditions (see Table 1). During the preshift phase of the experiment, infants in the three Single Talker conditions, were exposed to repetitions of an utterance of either /bʌg/ or /dʌg/ selected from one of the 12 different talkers. Half of the infants in each condition heard /dʌg/ and the other half, /bʌg/. To ensure that we had not selected the most discriminable pairs from one of our talkers, each infant in each control and experimental group was tested with tokens from a different talker. The infants in the control group heard the same token during both the habituation and postshift phase of the experiment. Infants in the Phonetic Change condition heard one of the two syllables from a particular talker during the habituation phase (e.g., /bʌg/) and the other one (e.g., /dʌg/) during the postshift phase. Infants in the Talker Change condition heard one syllable (e.g., /dʌg/ from Male#1) from a particular talker during the habituation phase and a phonetically identical syllable (i.e., /dʌg/ from Male#5) taken from a different talker of the same gender during the postshift phase. For half of the infants, the syllable type was /bʌg/ and for the other half, it was /dʌg/. Similarly, for half of the infants, the tokens were produced by female talkers, and for the other half, they were produced by male talkers.

The Multiple-Talker conditions were roughly parallel to the Single-Talker conditions. The Multiple-Talker control condition was identical to the Single-Talker control except that the tokens of a particular syllable type (e.g., /bʌg/) from all 12 talkers were presented in random order during each phase of the experiment. The Multiple-Talker Phonetic Change condition used tokens of a particular syllable (e.g., /dʌg/) from all 12 talkers during the habituation phase and were switched to the multiple tokens of the other syllable type (e.g., /bʌg/) during the postshift phase. Half of these infants heard /bʌg/ and half heard /dʌg/. Finally, the Mulitple-Talker Talker Change condition consisted of a habituation phase in which infants heard tokens of a particular syllable type (e.g., /bʌg/) spoken by a set of 6 different talkers (3 males and 3 females). Then, during the test phase, the infants were switched to an entirely new set of talkers uttering the same syllable. The particular talkers included in the habituation and postshift sets was varied randomly from infant to infant but always included 3 males and 3 females. Half of the infants heard utterances of /bʌg/ and the remainder, /dʌg/.

Insert Table 1 about here

Apparatus. A blind nipple was connected to a Grass PT5 volumetric pressure transducer, which in turn was coupled to a Grass (Model 7) polygraph. A Schmitt trigger provided a digital output of the criterial high-amplitude sucking responses. This output was relayed to a PDP 11/73 computer which recorded and saved the number of criterion responses on a minute by minute basis. In addition, it accessed the digitized syllables and controlled the presentation of the auditory stimuli at a level of 72 + 2 dB (C) SPL in response to criterion level sucking. The sounds were played out using a Kenwood (KA-3500) amplifier and a JBL (4310) loudspeaker. The computer was programmed to record the level of baseline responding, detect the attainment of the criterion for habituation, select the appropriate set of postshift stimuli, and terminate the experiment in the event that the criterion for habituation was achieved after 4 minutes during the postshift period.

Subjects. The subjects were 72 infants (36 males and 36 females) from the Eugene area with a mean age of 9.2 weeks. To obtain the 72 infants for this study, it was necessary to test 136 subjects. Subjects were excluded for the following reasons: crying (45%), falling asleep prior to shift (11%), repeatedly rejecting the pacifier (15.5%), ceasing to suck during the course of the experiment (e.g., 2 consecutive minutes of zero level responding) (11%), failure to achieve the habituation criterion within 24 minutes (12.5%) and miscellaneous (e.g., equipment failure, bowel movement, etc.) (5%).

Results

For purposes of statistical comparison, subjects' sucking rates were examined for four intervals: baseline minute, third minute before shift, average of minutes 1 and 2 before shift, and average of the first 2 minutes after shift. These data were then used to calculate difference scores for each of the following rate comparisons: (a) acquisition of the sucking response: third minute before shift - baseline; (b) habituation: third minute before shift - average of the last two minutes before shift; (c) release from habituation: average of the first two minutes after shift - average of the last two minutes before shift.[1]

---

[1] In addition, we also calculated a measure of the release from satiation for the full four minutes after shift (i.e., average of all four minutes after shift - average of last two minutes before shift). However, since the pattern of results with this measure was identical to that observed with the two-minute measure for all the experiments in the paper, we report only the two-minute measure since it is recognized in the literature as the more sensitiveof the two.

Table 1

Design of Experiment 1

Single Talker Conditions		
	Preshift Phase	Postshift Phase
Phonetic Change	bug(male#1)	dug(male#1)
Talker Change	bug(male#1)	bug(male#5)
Control	bug(female#4)	bug(female#4)
Multiple Talker Conditions		
	Preshift Phase	Postshift Phase
Phonetic Change	bug(f4),bug(m3),bug(f2), bug(m6),bug(f1),bug(m4), bug(m5),bug(f6),bug(f2),...	dug(f4),dug(m3),dug(f2), dug(m6),dug(f1),dug(m4), dug(m5),dug(f6),dug(f2),
Talker Change	bug(f4),bug(m3),bug(f2), bug(m6),bug(f1),bug(m4), bug(m4),bug(f1),bug(f2),...	bug(f3),bug(m2),bug(f6), bug(m5),bug(f5),bug(m1), bug(m1),bug(f6),bug(f5),
Control	bug(f4),bug(m3),bug(f2), bug(m6),bug(f1),bug(m4), bug(m5),bug(f6),bug(f2),...	bug(f4),bug(m3),bug(f2), bug(m6),bug(f1),bug(m4), bug(m5),bug(f6),bug(f2),

As is usually the case in studies employing the HAS procedure, subjects in all groups acquired the conditioned high-amplitude sucking response and attained the habituation criterion. Moreover, an ANOVA used to assess possible group differences during the preshift period revealed only the expected significant effect of minutes [$F(3,264) = 113.95, p < .0001$]. There was no evidence of any significant main effect for groups [$F(5,264) < 1.00$] or interaction of this variable with minutes [$F(15,264) < 1.00$].

The data concerning release from habituation during the postshift period are displayed in Figure 1. Randomization tests for independent samples (Siegel, 1956) were used to assess postshift sucking performance. The release from habituation scores of each experimental group were compared to its appropriate control group (i.e., the Single-Talker groups with the Single Talker control and the Multiple-Talker groups with the Multiple-Talker control). The results indicated that the infants showed significant ($p < .001$ or better) increases in sucking to the Phonetic Change in both the Single- and Multiple-Talker conditions, $t(22) = 4.09$ and 2.62, respectively. Thus, both groups discriminated the difference between /bʌg/ and /dʌg/.

Insert Figure 1 about here

The results from the Talker Change conditions presented a different pattern. Infants in the Single-Talker condition readily detected the Talker Change during the postshift period [$t(22) = 8.28, p < .001$]. This is an indication that even within the same gender, the differences in talkers' voices were highly discriminable for the infants. Nevertheless, there are apparently some limits on the ability of infants this age to encode information about talker identity, because infants in the Multiple-Talker Talker Change condition did not display evidence of discriminating the difference [$t(22) = 0.27$]. This finding replicates results reported for 7-month-old infants by Miller et al. (1982) in which they tried to train infants to respond to mixed groups of male and female talkers using a conditioned headturning procedure.

Thus far, the results indicate that, like their older counterparts (Kuhl, 1979; 1983), 2-month-olds are able to adjust to talker variability in detecting a phonetic contrast. However, to evaluate the consequences that adjusting to such variability might have on infants' processing of speech, several additional tests were conducted. First, we compared the postshift levels of responding by the infants in the Single- and Multiple-Talker Phonetic Change conditions using Randomization tests for independent samples. The two groups did not differ significantly with respect to this measure [$t(22) = 0.29$]. Next, we sought to determine whether a difference between infants in the Single- and Multiple-Talker conditions might be observed in the Time to Habituation measure used by Kuhl and Miller (1982). We collapsed across the three groups in both the Single- and Multiple-Talker conditions since the treatment in each of the groups was essentially the same for the preshift period. Subjects

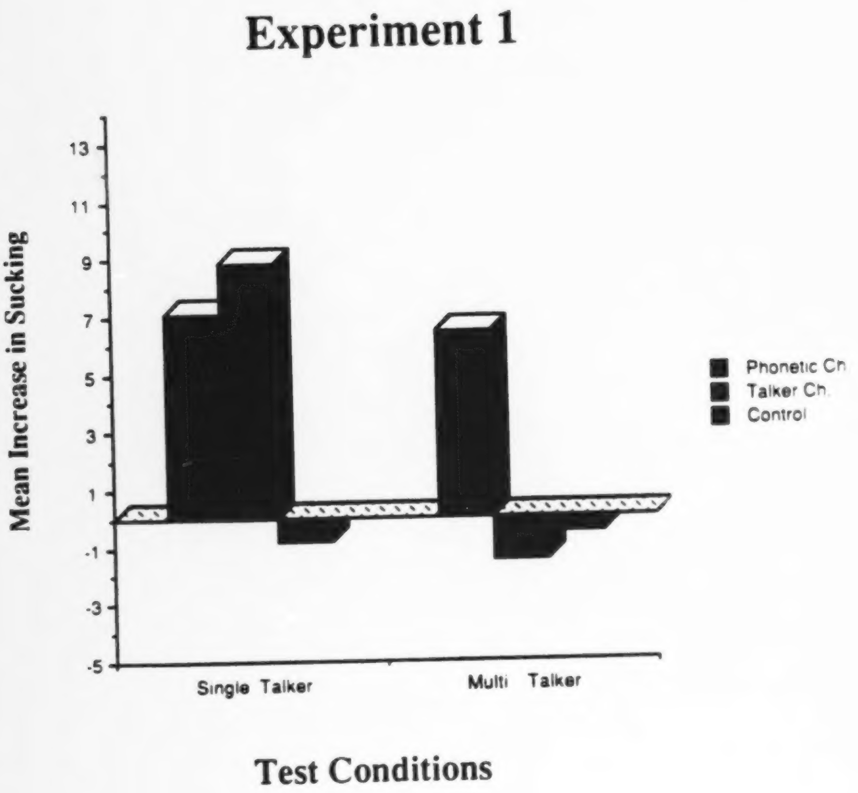


Figure 1. Mean change in postshift sucking for each of the Single Talker (left hand side) and Multiple Talker (right hand side) test conditions in Experiment 1. (The scores are determined by subtracting the average sucking rates from the last two preshift minutes from the average of the first two postshift minutes).

in the combined Multiple-Talker conditions took significantly longer to attain the habituation criterion than those in the combined Single-Talker conditions (9.61 and 7.58 minutes, respectively [$t(70) = 3.05, p > .005$]. Thus, the greater variety of tokens and/or their random patterning apparently sustains the interest of infants longer during the preshift phase of the experiment. A third analysis examined whether there was evidence that infants in the Multiple-Talker conditions might take longer to re-habituate to the stimuli during the postshift phase of the experiment. For this purpose, we used the measure employed by Bertoncini et al. (1988). For infants in each experimental group, we calculated the amount of time it took in the postshift period before the habituation criterion was achieved and/or the experiment was terminated.[2] The times to rehabituation scores are presented in Table 2. For the Phonetic Change conditions, no significant differences were observed between the Single- and Multiple-Talker groups [$t(22) = 0.22$]. However, when only a Talker Change was involved, the difference between the Single- and Multiple-Talker conditions was marginal, [$t(22) = 1.74, p < .10$], suggesting that re-habituation occurred more slowly for the Multiple-Talker group.

Discussion

What conclusions can be drawn from these results concerning the effects of talker variation on the perception of speech by 2-month-olds? First, it is clear that by this age infants already display some rudimentary form of perceptual normalization. Infants can detect a phonetic change between two stop consonants when as many as 12 talkers' voices vary irrelevantly. This result confirms and extends the findings reported for vowel and fricative contrasts reported by Kuhl (1979; 1983; Holmberg, Morgan & Kuhl, 1977) with 6-month-olds. Second, 2-month-olds are capable of perceiving differences between two different talkers' utterances of the same syllable, although there do seem to be limits to precisely what they can encode about talker differences. Thus, when listening to a set of different talkers utter the same syllable, they do not detect a change to a new set of talkers uttering the same syllable. Third, discrimination performance does not suffer significantly when tokens from multiple talkers are used as opposed to when only tokens from a single talker are used. Hence, unlike Kuhl & Miller's (1982) finding that irrelevant vowel variation significantly interfered with infants' ability to detect a pitch change, there is no evidence in the present study that talker variability interferes with infants' detection of a phonetic contrast. However, this is not to say that the infants' processing of speech is unaffected by talker variation. In fact, infants exposed to tokens from a variety of different talkers took significantly longer to habituate to syllables than infants who were exposed to tokens from a single talker.

Perhaps the most striking finding from this study is that infants listening to syllables produced by many different talkers did so well in detecting the phonetic change. Can it then be assumed that, in contrast to adults, lower level perceptual processes in infants are unaffected by talker variation? Such an assumption would be premature for a variety of reasons. First, we note that talker variation did have an effect on the time it took infants to habituate to the sounds. Hence, talker variability may have been affecting the development of a perceptual representation of the sounds. Second, the lower level perceptual effects caused by talker variation that have been reported for adults (Mullennix et al., 1989) are most evident when the stimulus conditions are less than optimal. For example, the longest effects occurred when the stimuli were degraded with noise. It is possible that under similar circumstances infants might also show deficits in discrimination performance in the presence of talker variation. Moreover, as noted earlier, encoding processes in memory are also affected in adults when talker variation is present (Martin et al., 1989). Given our finding of longer habituation times for the Multiple-Talker conditions, we wondered whether infants might also be affected by talker variation in the way that they encode and remember speech information. For this reason, we carried out another experiment.

---

[2] Only the data from the experimental groups are used in this calculation because responding in the control groups is already near floor levels and no new stimulation is introduced during the postshift period for these groups.

Experiment 2

The subject of how speech signals are encoded and remembered by infants has been discussed in the past. For example, in their study of vowel perception by infants, Swoboda, Morse and Leavitt (1976) noted that the likelihood that infants discriminated certain vowel contrasts appeared to be inversely related to the length of the interval between the last occurrence of the preshift stimulus and the first occurrence of the postshift stimulus when the HAS procedure was employed. Morse (1978) later suggested that manipulations of the preshift-postshift interval duration could provide a way of assessing memory effects in the HAS paradigm. In addition, discussions of the role that memory may play in discrimination performance have been raised in conjunction with studies that have used varied stimulus sets in place of a single stimulus during the habituation phase of the HAS procedure (e.g., Bertoncini et al., 1988; Jusczyk & Derrah, 1987; Kuhl & Miller, 1982; Miller & Eimas, 1979). Nevertheless, until very recently direct attempts to manipulate and assess memory factors in the HAS procedure have not been reported in the literature.

Clearly, information about the encoding process that infants use for speech is critical to understanding the growth and development of a lexicon in the native language. The infant must ultimately store some sort of acoustic-phonetic representation that will allow him or her to access the meanings of spoken words (see Jusczyk, 1985a; 1986; in press for further discussion of this point). In the present context, one can ask about the way in which talker variation might influence the encoding of speech sounds by infants. On the one hand, talker variation might be expected to interfere with encoding processes, as Martin et al. (1989) observed for adults. On the other hand, it might be argued that tokens from multiple talkers might permit infants to form a prototype that would actually facilitate the recognition of a syllable or word type. Thus, Grieser and Kuhl (1989) have recently reported evidence consistent with the view that 6-month-old infants may form prototypes for some speech sound categories and that "this may contribute to their seemingly efficient processing of speech information..." (p. 577). Indeed, one way of interpreting the lack of discrimination by infants in the Multiple Talker Talker Change condition in the previous experiment is that the infants formed a prototype for the syllable category that they were exposed to during the habituation phase and that the new instances that they heard during the postshift phase were simply treated as members of a familiar category.

The first step toward understanding the consequences of talker variation on encoding by infants is to devise a means for assessing their representation and memory of speech sounds. In addition to modifying the traditional HAS procedure by presenting a randomized set of sounds as in the previous experiment (see also Bertoncini et al., 1988; Jusczyk & Derrah, 1987; Kuhl & Miller, 1982), we also introduced another modification first employed by Jusczyk, Kennedy & Jusczyk (in preparation). Specifically, a 2-minute delay period filled with a slide presentation is introduced between the habituation and postshift phases of the HAS procedure. No auditory stimulation is present during this period. When the slide presentation is completed, the postshift period begins and the auditory stimulation resumes with either novel or familiar stimuli depending on whether an experimental or control condition is involved.

The basic issue is to determine whether talker variation affects infants' encoding of speech in long term memory. Consequently, we decided to compare performance under both single- and multiple-talker conditions. Four groups of infants were tested in conditions that paralleled the Phonetic Change and Control conditions of Experiment 1 (Single-Talker Phonetic Change and Control Conditions, and Multiple-Talker Phonetic Change and Control Conditions). If talker variation disrupts encoding, then discrimination performance in the Multiple-Talker condition should be worse than for the Single-Talker condition. On the other hand, if talker variation promotes the formation of prototypes, then performance may actually be better in the Multiple-Talker condition. Finally, in addition to these four groups, a fifth group, Single-Talker Talker Change Condition, was included in order to see whether infants might encode information about talker identity into their representations of syllables. To the extent that information about talker identity is stored, one would expect to find that infants would respond to the Talker Change after the delay interval.[3]

Method

Procedure. A modified version of the high-amplitude sucking procedure described in the previous experiment was used. The modification consisted of the insertion of a 2-minute delay interval between the habituation and postshift phases of the experiment. Upon the attainment of the habituation criterion, the computer beeped signaling to the experimenter in the control room to initiate the slide show. The fixation slide was extinguished and in its place, new slides were projected. The slides were a series of 24 colorful family vacation slides that were projected on the wall facing the infant in the test room. Each slide was shown for 5 sec. During the slide presentation, the experimenter in the test room continued to hold the pacifier in the infant's mouth although no auditory stimulation was presented. Following the 24th slide, the fixation slide was projected once again and auditory stimulation was available in response to criterion sucking. In all other respects, the procedure was identical to that used in Experiment 1. Extensive pilot testing by Jusczyk, Kennedy and Jusczyk (in preparation) determined the parameters for the memory delay interval. For example, the decision to keep the pacifier in place during the delay was made when it was determined that the removal and re-insertion of the pacifier during the delay interval led to spurious increases in sucking in the control and experimental groups. Similarly, the number of slides employed and their projection durations were optimal for maintaining the infants' attention.

---

[3] The parallel Talker Change Condition for Multiple-Talkers was not tested because of the failure of the infants in Experiment 1 to discriminate the difference even when no delay interval was employed.

Apparatus. The apparatus used was identical to that described for the previous experiment.

Stimuli. The same stimulus materials were used as in the previous experiment.

Design. Each infant was seen for one experimental session. Twelve subjects were assigned randomly to each of 5 test groups. Two of these groups employed tokens of /bʌg/ and /dʌg/ from all 12 talkers. For the Multiple-Talker Phonetic Change condition, randomly ordered tokens of one syllable type (/bʌg/ for half the infants, /dʌg/ for the other half) were presented during the habituation phase, and tokens of the other syllable type were played during the postshift phase. For the Multiple Talker Control condition, one of the two syllable types spoken by all 12 talkers was presented for both phases of the experiment. Two other groups heard tokens produced by a single talker for the entire test session (although the identity of the talker varied for each infant). For the Single-Talker Phonetic Change condition, one syllable (/bʌg/ for half the infants, /dʌg/ for the other half) was played during the habituation phase and the other syllable was played during the postshift phase. For the Single-Talker Control condition, one of these two syllables was presented for both phases. Finally, the Single-Talker Talker Change condition, employed tokens of the same syllable spoken by two different talkers of the same gender. During the habituation phase, the token from one talker was played and during the postshift phase, the token from the other talker was played. Once again, each infant heard a different pair of talkers. Half of the subjects heard a female pair and half heard a male pair. Similarly, half of the subjects listened to versions of /bʌg/ and the other half listened to versions of /dʌg/.

Subjects. The subjects were 60 infants (32 males and 28 females) from the Eugene area with a mean age of 7.4 weeks. To obtain the 60 infants for this study, it was necessary to test 121. Subjects were excluded for the following reasons: crying (51%), falling asleep prior to shift (16%), repeatedly rejecting the pacifier (18%), failure to achieve the habituation criterion within 24 minutes (11.5%), miscellaneous (experimenter error, parental interference) (3.5%).

Results

The data were analyzed as in the previous experiment. Difference scores were calculated for each subject to assess (a) acquisition of the sucking response, (b) habituation to the preshift stimuli, and (c) release from habituation during the first 2 minutes of the postshift period. As in the previous experiment, all groups acquired the conditioned response and habituated to the preshift stimuli. Moreover, an ANOVA used to assess possible group differences during the preshift period revealed only the expected significant effect of minutes $[F(3,220) = 194.93, p < .0001]$. Neither the main effect for groups $[F(4,220) = 2.036, p = .09]$ nor the interaction of this variable with minutes $[F(12,220) = 0.648, p = .80]$ was statistically significant.

The data on release from habituation are shown in Figure 2. Randomization tests for independent samples were again used to assess postshift sucking performance. In contrast to the previous experiment, a difference emerged in the way in which infants in the Single- and Multiple-Talker conditions responded to the Phonetic Change after the delay period. In particular, only in the Single-Talker condition did the Phonetic Change group show a significant increase in sucking relative to the Control group during the postshift period $[t(22) = 2.11, p = .046]$. Not only was the difference between the Phonetic Change and Control groups not significant for Multiple Talker conditions $[t(22) = -0.29]$ but it was even in the wrong direction. Thus, the presence of talker variation affects encoding of speech sounds in memory by young infants.

Insert Figure 2 about here

Performance in the Single-Talker Talker Change group was also different than the results observed in Experiment 1. When compared to the Single Talker Control group, infants in the Talker Change group did not exhibit a significant increase in postshift sucking $[t(22) = 0.89, p = .38]$. This suggests that talker identity may not have a high priority with respect to the kind of information that infants encode and/or retrieve about speech sounds.

As in the previous experiment, we also examined the impact of talker variation on the time to achieve the habituation criterion in both the preshift and postshift phases of the experiment. For the preshift phase, we collapsed across all the Single-Talker groups and across both Multiple-Talker groups since the stimulus presentation was the same for this

Experiment 2

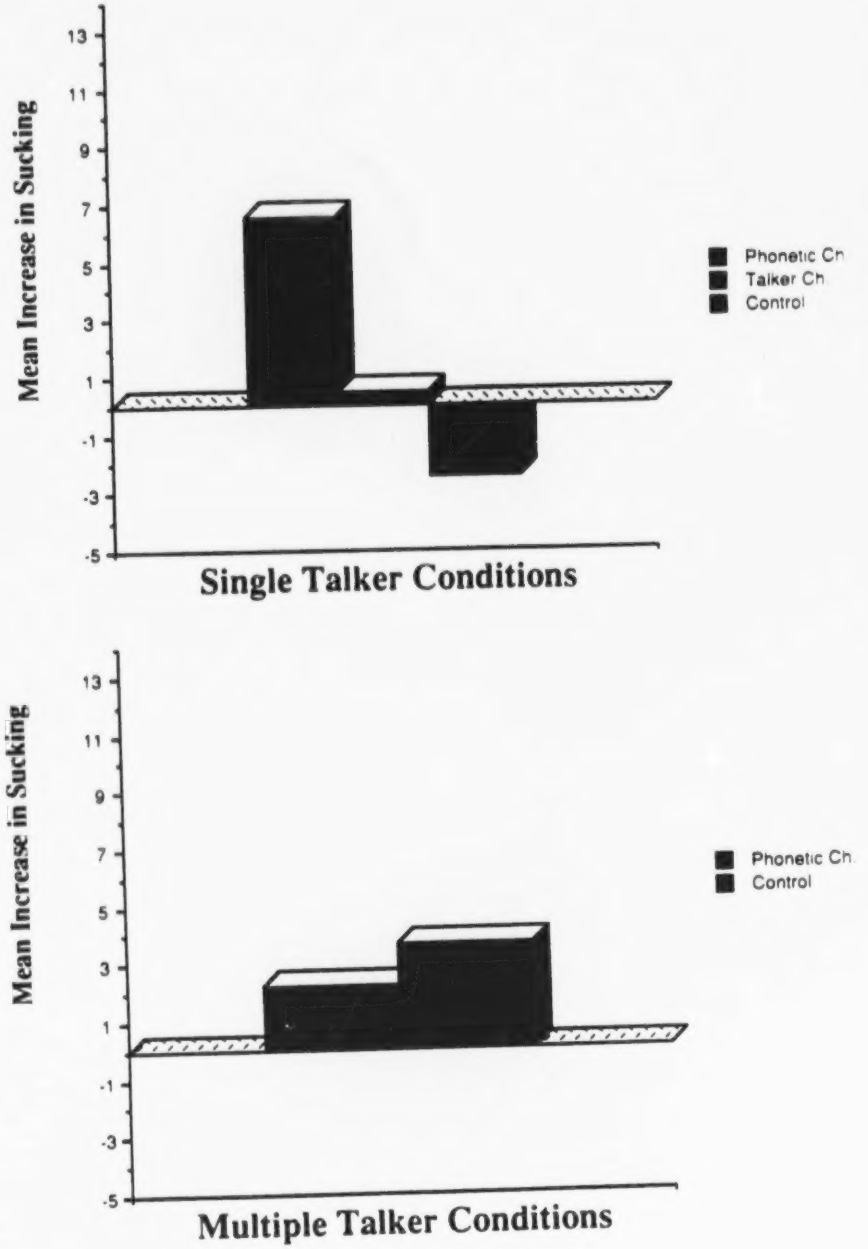


Figure 2. (a) Shows the mean change in postshift sucking for each of the Single Talker conditions after the 2-minute delay period in Experiment 2. (b) Shows the comparable results for the Multiple Talker conditions.

period. Once again, there was evidence of significantly longer times to habituation [$t(58) = 2.53, p < .02$] for the Multiple-Talker group (11.74 minutes) than for the Single-Talker group (9.22 minutes). To evaluate re-habituation during the postshift period, the comparable groups were the Single-Talker Phonetic Change (5.08 minutes) and Multiple Talker Phonetic Change (5.58 minutes) groups. There was no evidence that these groups differed significantly on this measure [$t(22) = 0.76$]. Hence, talker variation appears to have affected the time to habituation only during the preshift phase of the experiment.

Discussion

Two-month-old infants are able to retain acoustic-phonetic information for a delay period of 2-minutes. This is evident in the performance of infants in the Single-Talker Phonetic Change group to detect the difference between the stimuli played during the preshift and postshift periods. Nevertheless, it is also clear that talker variation disrupts encoding and/or retrieval processes in infants this age. Thus, infants in the Multiple-Talker Phonetic Change group did not detect the difference between the preshift and postshift stimuli. The locus of this effect appears to be in the encoding processes associated with long-term memory. In the previous experiment without the delay, infants in the multiple talker group were able to perceive the very same phonetic change. Hence, as in adults (Martin et al., 1989), we find evidence that when infants are exposed to different talkers, this stimulus variability can disrupt encoding processes.

Given the kind of experience that infants received in the present experiment, it is also clear that exposure to different talkers uttering the same syllable did not facilitate the formation of a prototype. Rather, talker variation appeared to interfere with the way in which infants encoded speech information. This is shown not only in the failure of the infants in the Multiple-Talker Phonetic Change group to discriminate the contrast, but also by the fact that they took much longer to habituate to the syllables in the first place. One possible explanation of the difficulty is that the infants were trying to encode the syllables individually using talker specific cues. However, this explanation seems unlikely in view of the fact that infants in the Single-Talker Talker Change group gave no evidence of retaining information about talker identity over the delay period, despite the fact that their counterparts in Experiment 1 did detect such a change in the absence of any delay.

An alternative explanation of the present results is that the pattern observed here is not the result of talker variation, *per se*, but is due to the presence of multiple tokens in the familiarization phase combined with the delay in testing. Were this explanation correct, then one would expect that whenever multiple tokens are used during the preshift phase and testing is delayed, infants should fail to detect the presence of new items in the test phase. However, Jusczyk et al. (in preparation) used a series of phonetically distinct syllables in the preshift phase of their experiment and found that 2-month-olds did detect phonetic changes after a 2-minute delay in testing. Therefore, the decrements in performance observed in the present study had more to do with the kind of information that was varying (talkers' voices) than the mere fact that something was varying. However, to explore further the consequences that talker variation has on speech processing by infants, we sought to determine whether talker variation always disrupts memory for speech sounds. To evaluate this possibility, we decided to investigate whether talker variation affects the ability to remember a very salient distinction, viz., a change between male and female voices.

Experiment 3

Miller et al. (1982) established that 7-month-old infants could readily learn to categorize male and female voices. Moreover, they demonstrated that infants' success on the task was not attributable to use of fundamental frequency of the voices (males have generally lower fundamental frequencies than females) to distinguish the categories. Thus, there is some reason to believe that differences between male and female voices may be quite salient for infants. Accordingly, we examined whether infants exposed to a variety of talkers from one gender would retain this information over a 2-minute delay interval so as to notice a change to talkers of the opposite gender. Because the main objective of the study was to determine whether talker variation disrupts the memory for any sort of speech contrast, we tested infants only on multiple talker stimuli. Hence, the present experiment had only two test groups. One group was a Talker Gender Change group in which infants were exposed to utterances of a particular syllable by 6 talkers of one gender during the preshift period, and to utterances of the same syllable by 6 talkers of the opposite gender in the postshift period after a 2-minute delay. The other test group was a Gender Control group in which infants were exposed to utterances of a particular syllable by 6 talkers of the same gender throughout the entire test session.

Method

Procedure and Apparatus. The procedure and apparatus were identical to that of Experiment 2.

Stimuli. The same stimulus materials were used as in the previous two experiments.

Design. Each infant was seen for one experimental session. Twelve subjects were assigned to each of two test groups. Infants in the Talker Gender Change group heard randomly ordered tokens of one syllable type (half heard /bʌg/, half heard /dʌg/) produced by either 6 male or 6 female talkers during the preshift phase of the experiment. During the postshift period which began after a 2-minute delay interval, the infants heard utterances of the same syllable type produced by 6 talkers of the opposite gender. Infants in the Gender Control group, were treated in the same way for the preshift period, but during the postshift period they continued to hear the same utterances that they had heard prior to the 2-minute delay. For the infants in this group, half of them heard utterances from females and half heard utterances from males. Similarly, half heard /bʌg/ and half heard /dʌg/.

Subjects. The subjects were 24 infants (11 males and 13 females) from the Eugene area with a mean age of 8.7 weeks. In order to obtain the 24 infants for the study, it was necessary to test 47. Subjects were excluded for the following reasons: crying (43.5%), falling asleep prior to shift (17.5%), repeatedly rejecting the pacifier (30%), and failure to attain the habituation criterion within 24 minutes (9%).

Results

The data were analyzed as in the previous two experiments. Difference scores were calculated for each subject to assess (a) acquisition of the sucking response, (b) habituation to the preshift stimuli, and (c) release from habituation during the first 2 minutes of the postshift period. As in the previous experiments, both groups acquired the conditioned response and habituated to the preshift stimulus. Moreover an ANOVA used to assess possible group differences during the preshift period revealed only the anticipated significant effect of minutes $[F(3,88) = 50.69, p < .0001]$. Neither the main effect for groups $[F(1,88) = 1.95, p = .166]$ nor the interaction of this variable with minutes $[F(3,88) = 0.35, p = .787]$ was statistically significant.

The data on release from habituation are shown in Figure 3. Randomization tests for independent samples, used to assess postshift sucking performance, indicated a significant difference between the Talker Gender Change group and the Gender Control group $[t(22) = 3.923, p = 001]$. Therefore, despite the presence talker variation, 2-month-olds were able to retain information about the gender of the talkers over a 2-minute delay interval.

Insert Figure 3 about here

Discussion

The presence of some forms of talker variation does not entirely disrupt the retention of all information about speech by 2-month-olds. In the present case, infants did appear to

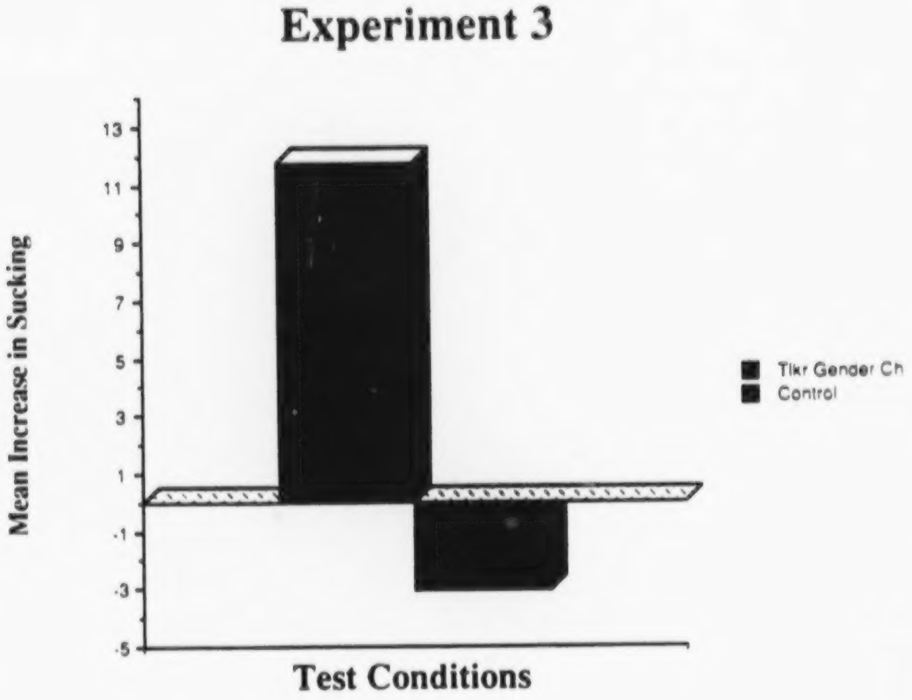


Figure 3. Shows the mean change in postshift sucking for each of the Single Talker conditions after the 2-minute delay period in the Talker Gender Change and Control conditions of Experiment 3.

experience a difficulty in detecting the change from talkers of one gender to talkers of the opposite gender. Two factors may have played a role in setting the present situation apart from the one in Experiment 2. First, there was obviously greater homogeneity in the groups of talkers in the present experiment than in the previous one. Second, the nature of the contrast itself, between male and female voices, may have been more discriminable than the phonetic contrast between two stop consonants that was tested in the previous experiment. Indeed, Kuhl and Miller (1982) offered a similar argument to explain why their infants were able to detect one type of change (a vowel distinction) but not another (a pitch contrast).[4]

The results of the present experiment also replicate the basic findings of Miller et al. (1982) and extend them to a younger age group. Hence, 2-month-olds are able to distinguish between male and female speaking voices. Moreover, they are able to retain information about the gender of the talker for at least a short 2-minute interval. The extent to which information about the talker's gender might be retained for longer intervals has yet to be determined.

General Discussion

The present study demonstrates that infants as young as 2-months of age have some capacity to cope with talker variation during speech perception. This finding replicates those reported by Kuhl (1979; 1983) with 6-month-old infants. However, the present study also shows that dealing with such variation also carries some costs with respect to the way that speech is processed. For instance, infants took longer to habituate to repetitions of a particular syllable when talker variation was present. Moreover, the presence of talker variation apparently disrupted infants' encoding of speech information in a way that prevented them from detecting a phonetic change which occurred over a short delay interval. Thus, the consequences of talker variation on infants' ability to process speech information appear to occur chiefly in the way that memory for the material is affected. In this respect, our results are similar to ones reported for adults. Specifically, in the absence of any noise induced degradation of the speech signal, there is little evidence that perceptual processes related to the identification of items are disrupted significantly in adults (Mullennix et al., 1989), whereas the mere presence of talker variation is sufficient to adversely affect processes associated with the retention of speech information (Martin et al., 1989). This further similarity in the way in which perceptual normalization processes operate in both infants and adults is certainly at least consistent with Miller's (1987) contention that the mechanisms underlying these processes may be innately prewired. Of course, there may be other aspects of perceptual normalization, of which we are yet unaware, that are either incomplete at this age or require further experience with a native language.

---

[4] However, note that in Kuhl and Miller's study, disruptions caused by irrelevant variation on a dimension had an impact on the immediate detection of the contrast on the critical dimension. Whereas, in the present study, irrelevant talker variation only had consequences for the discriminability of a phonetic contrast when a delay occurred in testing, as in Experiment 2. The same phonetic change was detected in Experiment 1 when there was no delay between the preshift and postshift periods.

Another important domain that the present results bear upon concerns the kind of information that infants retain about speech sounds. By comparing how infants perform on the same contrasts under conditions of delay and no-delay, we were able to gain some appreciation of the kind of information that is most likely to be retained upon hearing speech. As noted earlier, previous research by Jusczyk et al. (in preparation) demonstrated that infants are able to retain information about the phonetic features of syllables for a short delay period. The performance of the infants in the Single Talker Phonetic Change condition of Experiment 2 replicated this basic finding. Hence, information relevant to the phonetic coding of speech sounds is one type of information that infants are likely to retain. However, it is worth noting that retention of this type of information is adversely affected by the presence of talker variation. One potential explanation of the latter result is that infants are also trying to encode information about talker identity and that this somehow interferes with their storage of information concerning phonetic features. Yet this sort of explanation is apparently ruled out when we consider what happened to the retention of information about talker identity. When infants were listening to a token of a particular syllable produced by a single talker before the delay and were switched to an utterance of the same syllable produced by a new talker, they did not appear to retain information about talker identity. Moreover, this failure to retain information about talker identity was not due to an inability to detect the difference in talkers' voices because infants who heard the same pairs of syllables without the delays did discriminate them. Still, it would be too strong to claim that infants encode information about phonetic features but not about talker characteristics, as is evident by the performance of the infants in the Talker Gender Change condition of Experiment 3. Apparently, information about some talker characteristics, in this case gender, may be encoded. Indeed, it is tempting to speculate how infants would perform if talker identity were made particularly salient for infants by including some very familiar voices as opposed to a set of total strangers as in the present case. Perhaps one would find that under such circumstances, infants do retain information about particular talker characteristics.

Information about what information infants retain from speech sounds is certainly critical in understanding how a lexicon develops that serves speech recognition in a native language. Recognizing a word in fluent speech requires that elements in the sound stream activate the correct stored meaning. It is not obvious how this could be accomplished in the absence of some stored representation of the sound pattern of the word. One of the long term goals of research on infant speech perception as it relates to the development of the lexicon is to determine the kind of information that goes into the infant's representation of the acoustic-phonetic characteristics of words (see Jusczyk, in press, for further discussion of this point). If information about the identity of a talker figures into the representation then this has certain consequences for models of word recognition. In fact, such a result would be difficult to handle for models that postulate the storage of some prototypical representation of the acoustic-phonetic characteristics of lexical items because differences among pronunciations of the same word by different talkers is just the kind of information that a prototype might be expected to exclude. Instead, exemplar models, ones that postulate that listeners store traces of particular utterances that they hear, would be favored by results suggesting that talker characteristics are retained in representations. In addition to providing a more straightforward account of why recognition of previously encountered instances from a category tends to be better, these sorts of models account for the same range of facts as prototype models (see Hintzman, 1986 for an interesting discussion of this point).

As noted earlier, Grieser and Kuhl (1989) have examined the issue of whether 6-month-old infants may form prototypes for certain vowel categories. In their study, they compared generalization performance to novel instances from a category after exposure to good (prototypical) and poor exemplars from the category. Performance was significantly better in the case of exposure to the good exemplars. Grieser and Kuhl concluded that their results are consistent with a view that "holds that human infants organize vowel categories around prototypes" (p.577). We concur with this conclusion, but we also believe that some of the present findings should be taken into account in thinking about this issue. For example, although the finding that information about talker identity is not retained during a 2-minute delay interval is consistent with the view that speech information is encoded in the form of a prototype of the acoustic-phonetic properties of speech, what are we to make of the finding that information about talker gender is retained for this interval? Also, under normal circumstances one would expect that repeated exposure to a diverse set of exemplars from a category would make it more likely the detection of a change to a new category more likely than repeated exposure to a single instance from the category. Yet, precisely the opposite occurred in the present experiment. Clearly, it is premature to take a firm stand as to whether an exemplar or a prototype model best describes the way in which infants encode speech information. Moreover, it is certainly not our intention to attack Grieser and Kuhl's idea that a prototype description may provide the best account of the way speech sounds are represented by infants. Our point is only that any decision in favor of one or the other type of model will be possible only after considering a broad range of facts including the effects that talker variation has on the way in which speech sounds are recognized.

References

Allard, F. & Henderson, L. (1975). Physical and name codes in auditory memory: The pursuit of an analogy. *Quarterly Journal of Experimental Psychology*, **28**, 475–482.

Aslin, R.N. (1987). Visual and auditory development in infancy. In J.D. Osofsky (Ed.), *Handbook of infant development* (2nd Edition, pp. 5-97). New York: Wiley.

Aslin, R.N., Pisoni, D.B., & Jusczyk, P.W. (1983). Auditory development and speech perception in infancy. In M. Haith and J. Campos (Eds.), *Handbook of child psychology: vol.2. Infancy and developmental psychobiology.* (pp. 573-687). New York: Wiley.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P.W., Kennedy, L.J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, **117**, 21-33.

Bladon, R.A., Henton, C.G., & Pickering, J.B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication*, **4**, 59-69.

Carrell, T.D., Smith, L.B. & Pisoni, D.B. (1981). Some perceptual dependencies in speeded classification of vowel color and pitch. *Perception and Psychophysics*, **29**, 1–10.

Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory for a speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 1-7.

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.

Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, **29**, 655.

DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, **208**, 1174-1176.

Dechovitz, D. (1977). Information conveyed by vowels: a confirmation. *Haskins Laboratories Status Report on Speech Research, SR-51/52*, 213-219.

Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253-261.

Eimas, P.D. (1982). Speech perception: A view of the initial state and perceptual mechanisms. In J. Mehler, M. Garrett, & E.C.T. Walker (Eds.) *Perspectives on mental representation: Experimental and theoretical studies of cognitive processes and capacities.* (pp. 339-360). Hillsdale, NJ: Erlbaum.

Eimas, P.D., Siqueland, E.R., Jusczyk, P. & Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303-306.

Fourcin, A.J. (1968). Speech-source interference. *IEEE Transactions on Audio and Electroacoustics*, **ACC-16**, 65-67.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, **ACC-16**, 78-80.

Grieser, D., & Kuhl, P.K. (1989). The categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, **25**, 577-588.

Hintzman, D.L. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Holmberg, T.L., Morgan, K.A., & Kuhl, P.K. (1977). Speech perception in early infancy: Discrimination of fricative consonants. Paper presented at the meeting of the Acoustical Society of America, Miami Beach, Florida. December.

Huttenlocher, J. (1974). The origins of language comprehension. In R.L. Solso (Ed.) *Theories in cognitive psychology*. New York:Wiley.

Jusczyk, P.W. (1981). Infant speech perception: A critical appraisal. In P.D. Eimas & J.L. Miller (Eds.) *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Jusczyk, P.W. (1985a). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.) *Neonate cognition: Beyond the blooming, buzzing confusion* (pp. 199-229). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W. (1985b). The high-amplitude sucking procedure as a methodological tool in speech perception research. In G. Gottlieb & N.A. Krasnegor (Eds.) *Infant methodology*. Norwood, NJ: Ablex.

Jusczyk, P.W. (1986). Towards a model for the development of speech perception. In J. Perkell & D.H. Klatt (Eds.) *Invariance and variability in speech processes* (pp. 1-19). Hillsdale, NJ: Erlbaum.

Jusczyk, P.W. (in press). Developing phonological categories from the speech signal. In C.A. Ferguson & C. Stoel-Gammon (Eds.) *Phonological development*. Baltimore, MD: York Press.

Jusczyk, P.W., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L.J., & Mehler, J. (in press). The role of attention in speech perception by infants. *Cognitive Development*.

Jusczyk, P.W. & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, **23**, 648-654.

Jusczyk, P.W., Kennedy, L.J. & Jusczyk, A.M. (in preparation) Young infants' memory for information in speech syllables.

Kaplan, E.L. (1969). *The role of intonation in the acquisition of language.* Unpublished Ph.D. dissertation, Cornell University, Ithaca, N.Y.

Kuhl, P.K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, **66**, 1668-1679.

Kuhl, P.K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, **6**, 263-285.

Kuhl, P.K. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (Eds.) *Handbook of infant perception, Vol. 2* (pp. 275-381). New York: Academic.

Kuhl, P.K. & Miller, J.D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics*, **31**, 279-292.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **15**, 676-684.

Mehler, J., Bertoncini, J., Barriere, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, **7**, 491-497.

Miller, C.L., Younger, B.A., & Morse, P.A. (1982). Categorization of male and female voices in infancy. *Infant Behavior and Development*, **5**, 143-159.

Miller, J.L. (1987). Mandatory processing in speech perception. In J.L. Garfield (Ed.) *Modularity in Knowledge and natural-language understanding.* Cambridge, MA: MIT Press.

Miller, J.L. & Eimas, P.D. (1979). Organization in infant speech perception. *Canadian Journal of Psychology*, **33**, 353-367.

Mills, M. & Meluish, E. (1974). Recognition of the mother's voice in early infancy. *Nature*, **252**, 123-124.

Morse, P.A. (1978). Infant speech perception: Origins, processes and Alpha Centauri. In F. Minifie and L.L. Lloyd (Eds.) *Communicative and cognitive abilities: Early behavioral assessment.* Baltimore: University Park Press.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Neary, T.M. (1978). *Phonetic feature systems for vowels.* Indiana University Linguistics Club, Bloomington, IN.

Rand, T.C. (1971). Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research, SR-25/26*, 141-146.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Siqueland, E.R. & DeLucia, C.A. (1969). Visual reinforcement of non-nutritive sucking in human infants. *Science*, **165**, 1144-1146.

Summerfield, Q. (1975). Acoustic and phonetic components of the influence of voice changes and identification times for CVC syllables. *Report of Speech Research in Progress, 2(4)*, The Queen's University of Belfast, Belfast, Ireland.

Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalisation as demonstrated by reaction times. *Report of Speech Research in Progress, No. 2*, The Queen's University of Belfast, Belfast, Ireland.

Swoboda, P., Morse, P.A., & Leavitt, L.A. (1976). Continuous vowel discrimination in normal and at-risk infants. *Child Development*, **47**, 459-465.

Syrdal, A.K., & Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086-1100.

Turnure, C. (1971). Response to the voice of mother and stranger by babies in the first year. *Developmental Psychology*, **4**, 182-190.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976) What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, **60**, 198-212.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Neighborhood Density Effects for High Frequency Words: Evidence for Activation-based Models of Word Recognition[1]

Stephen D. Goldinger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This research was supported by NIDCD Research Grant R01 DC00111-13 to Indiana University, Bloomington, IN. I thank David Pisoni and Paul Luce for helpful comments and criticisms, and Denise Beike for her help in recording stimulus materials.

Abstract

Two experiments investigated the effects of lexical neighborhood density on the recognition of spoken words. All words employed in both experiments were the highest frequency words of their respective neighborhoods. According to frequency-ordered search-based models of word recognition (e.g., Forster, 1976), no effects of neighborhood size should be observed for the highest frequency word of any particular search set. Contrary to the search models' prediction, in both lexical decision and naming experiments, recognition of words from sparse neighborhoods was consistently faster than recognition of words from dense neighborhoods. The results are discussed in terms of both serial- and parallel-search models of word recognition.

Neighborhood Density Effects for High Frequency Words: Evidence for Activation-based Models of Word Recognition

Among topics of research in the field of cognitive psychology, few have received as much attention as the recognition of printed and spoken words. Indeed, although the predominant research efforts have been dedicated to investigating the perception of printed words, the same critical, modality-independent questions regarding the nature of the mental lexicon and its associated retrieval processes have remained central to all investigations and theories of word recognition for over two decades (see, e.g., Adams, 1979; Tyler & Frauenfelder, 1987, for review). The popularity and importance of word recognition research may be primarily attributed to the impressive speed of word recognition and lexical access relative to the size of the lexicon (Forster & Bednall, 1976; Marslen-Wilson & Welsh, 1978). Estimates of the number of words resident in the mental lexicon of the average adult vary from as few as 50,000 to as many as 250,000 (Seashore & Eckerson, 1940), yet recognition of any given word may occur in less than a quarter of a second. Regardless of the actual size of any given individual's lexicon, such high estimates attest to the impressive perceptual abilities that "word recognition" entails. Accordingly, the primary task of researchers in word recognition has been to derive models that effectively communicate the efficiency of the processes involved, while simultaneously respecting the theoretical constraints provided by well-known phenomena, such as frequency effects and context effects.

The present studies were conducted to add to the growing body of data concerning the effects of *lexical neighborhood* characteristics on the recognition of spoken words, and to evaluate several broad classes of models for their adequacy to explain neighborhood effects. A lexical neighborhood may be defined (for spoken words) as a collection of words that sound similar to a given word. Typically, neighbor relationships among spoken words have been determined by use of either acoustic confusion data, or by phoneme-substitution algorithms (see, e.g., Luce, 1986a). Similarly, for written words, the majority of research conducted on neighborhood effects has used the letter-substitution *N* metric developed by Coltheart, Davelaar, Jonasson, and Besner (1977).

Investigations into the effects of neighborhood size and structure on the recognition of their constituent words may provide deeper insight into the nature of the word recognition process. Since the process of word recognition primarily entails the resolution of one meaningful stimulus pattern from a vast pool of potential patterns, it is obviously germane to study the effects of similarity relations among words in memory on the efficiency of recognition. The investigation of neighborhood effects takes on even more importance when different classes of models are compared. A recent example of the potency of neighborhood effects for comparing across classes of models is provided by Andrews' (1989) experiments on the effects of orthographic neighborhoods on the recognition of visually-presented words.

Andrews' (1989) experiments:

Andrews (1989) recently reported the results of four experiments investigating the effects of orthographic neighborhood size on the recognition of low and high frequency words. In experiments using both the lexical decision and naming paradigms, larger neighborhoods were shown to *facilitate* the recognition of words, especially low frequency words. The beneficial effects of orthographic neighbors on word recognition was shown to be a true recognition effect, not explainable by recourse to the decision requirements of the lexical decision task (Balota & Chumbley, 1984) or the pronunciation requirements of the naming task (Balota & Chumbley, 1985).

Andrews conducted experiments to allow comparison of two general classes of word recognition models— activation-based and search-based models. Activation-based models of word recognition, such as the interactive activation model (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) are capable of explaining the beneficial effects of large neighborhoods via the system of positive feedback loops between the lexical and sublexical levels of units assumed in the model's architecture. As more potential word candidates are activated by a stimulus input word, the constituent segments and features of the word are quickly resolved by virtue of the increasing positive feedback provided by the lexical level. This "gang effect" should produce greater support for words from dense neighborhoods than for words from sparse neighborhoods, so the interactive activation model predicts greater facilitatory effects of larger neighborhoods. Unfortunately, as Andrews points out, the interactive activation model is equally capable of explaining precisely the opposite effect as well. Because McClelland and Rumelhart's model incorporates connections for intra-level inhibition as well as inter-level excitation, there is an appropriate mechanism available to account for *inhibitory* effects of larger neighborhoods as well. If it is assumed that the inhibition produced among nodes at the lexical level is substantially greater than the excitation produced between levels, the interactive activation model predicts inhibition from neighbors. Without principled selections of key parameters, therefore, the interactive activation model is not testable by data reporting only the effects of neighborhood size (Andrews, 1989).

Although the overly-powerful nature of the interactive activation model makes direct tests of the model difficult, it does not necessarily devalue the class of activation-based models as a whole. Andrews demonstrates this by comparing the noncommital "predictions" of the interactive activation model with stronger predictions derived from search-based models, particularly Forster's (1976) search model and the activation-verification model discussed by Becker (1976) and by Paap, Newsome, McDonald, and Schvaneveldt (1982). Fortunately, unlike the interactive activation model, search models *do* generate explicit qualitative predictions regarding the effects of large neighborhoods on the recognition of their constituent words. Unfortunately, for the case of visual word recognition, the predictions these models yield contradict the findings Andrews reported. The reason for this failure lies in the search models' treatment of word frequency; both Forster's search model and the activation-verification model explain word frequency effects by assuming a *frequency-ordered search* through some appropriate subset of the lexicon derived from gross sensory analysis of the input. Given the assumption of frequency-ordered search, it is clear that the only prediction search models make regarding neighborhood effects is that larger neighborhoods should tend to inhibit word recognition. Because the search process proceeds serially from the most frequent word in the neighborhood to the eventual target word, the more neighbors there are in the total search set, the longer it should take to reach the word in question. This effect should be most pronounced for low frequency words, since larger neighborhoods will tend to move low frequency words farther and farther down in the list to be searched. The effects of neighborhood size reported by Andrews, however, were that larger neighborhoods tend to *facilitate* the recognition of their constituent words, and the effect was *most* robust for low frequency words, in contrast to the prediction of the search-based models.[1]

Luce's (1986a) experiments:

For visual word recognition, although activation-based models cannot actually *predict* the facilitatory effects of large orthographic neighborhoods, it is clear that they do not *preclude* the effects, as do the search-based models. Unfortunately, in the case of spoken word recognition, the evidence is not so definitive. Unlike the findings reported by Andrews regarding the beneficial effects of neighbors for printed words, recent findings reported by Luce (1986a) and by Goldinger, Luce, and Pisoni (1989) show that the recognition of spoken words is *inhibited* by the presence of many neighbors.

Luce (1986a; see also Luce, Pisoni, & Goldinger, in press) has reported findings on the effects of lexical neighborhood characteristics on the recognition of spoken words. In experiments using perceptual identification of words in noise, auditory lexical decision, and auditory word naming, two effects were consistently observed: (1) words from sparse neighborhoods were recognized more quickly and accurately than words from dense neighborhoods, and (2) words with few higher-frequency neighbors were recognized more quickly and accurately than words with many higher-frequency neighbors. Both of these findings reveal the competitive and inhibitory nature of lexical neighborhoods on the recognition of spoken words. Similar findings were obtained by Goldinger, Luce, and Pisoni (1989).

Comparing the modality-specific effects of lexical neighborhoods, we see that dense neighborhoods are beneficial to the recognition of printed words, but are detrimental to the recognition of spoken words. This conflicting nature of neighborhood density effects on word recognition calls the generality of Andrews' (1989) findings into question. Although it is easily demonstrated that search-based models cannot account for the facilitatory neighborhood effects observed for visual word recognition, models such as Forster's (1976) search model *can* account for the inhibitory effects of dense neighborhoods observed for spoken words. As Andrews points out, inhibition from dense neighborhoods is a natural prediction of the frequency-ordered search process, since words from dense neighborhoods are likely to have more high frequency neighbors that will delay the search process. Therefore, although it has been argued that activation-based models are the most appropriate models for visual word recognition, the possibilities remain open for spoken word recognition.

---

[1] It should be noted that Andrews' results are subject to a second possible interpretation that does not entail neighborhood density, per se. It is well-established in the visual word recognition literature that both letter-positional frequency and bigram frequency affect word recognition speed (Mason, 1975; Massaro, Venezky, & Taylor, 1979). It is easily shown that as orthographic neighborhood size increases, relative positional and bigram frequencies of the letters constituting the neighborhood increase as well. Therefore, the beneficial effects of large neighborhoods *could* merely represent a frequency effect.

The present study is intended as a spoken word recognition analog to Andrews' (1989) visual word recognition experiments. Accordingly, the experiments reported here examined the recognition of spoken words from dense and sparse neighborhoods. However, unlike Andrews' experiments, in which the *low* frequency neighbors provided the critical test case, in the present experiments it is the recognition of *high* frequency neighbors that can provide diagnostic power. Specifically, the present experiments investigated the recognition of the highest frequency words from dense and sparse neighborhoods.

The logic behind investigating the recognition of these "very high frequency" words is derived from a central prediction of all search-based models. In models such as Forster's search model or the activation-verification model, the highest frequency word of a pool of candidates is always the first word checked for possible recognition. If there is a sufficient match between this initial candidate and the stimulus input, search is terminated and the word is recognized (Forster & Bednall, 1976). This interpretation of frequency effects in search-based models provides a convenient analytic tool for the investigation of neighborhood effects. Although search models can predict inhibitory effects of neighbors on word recognition, the neighborhood density effect is only an indirect by-product of the frequency-ordered search process. Considering the recognition of only the highest frequency words from each neighborhood, it becomes apparent that the search models predict *no differences* in recognition times for words from dense and sparse neighborhoods. If the highest frequency word from each neighborhood is always the first candidate checked, and if search terminates upon recognition, then the recognition speeds for the highest frequency words from both dense and sparse neighborhoods should be equivalent.

Thus, search-based models logically preclude neighborhood density effects for the highest frequency word of any kind of neighborhood. Unfortunately, as was the case for visual word recognition, qualitative predictions are not derived as easily for activation-based models, such as the interactive activation model. For the interactive activation model, predictions regarding the net effects of neighborhood density on the highest frequency neighbor depend on complex interactions of several unspecified parameters (see Andrews, 1989, for discussion). For any given word, the overall recognition speed is a combined product of the word's frequency, number of neighbors, absolute similarity to neighbors, and the particular values assigned to the excitatory and inhibitory connections in the system. Because of the inherent complexity of such highly interactive models, the present experiments cannot directly address any specific predictions of the genre, per se. Nevertheless, the two most common classes of models in the word recognition literature are search-based and activation-based models. By carefully scrutinizing the more testable class, we may at least obtain indirect support for the other.

In summary, the present investigation examined subjects' speed of recognition for words that are the highest frequency words of their respective neighborhoods. The question of major interest is whether or not variations of neighborhood density affect such privileged words. While predictions are not easily derived for highly interactive models, such as the interactive activation model, search-based models clearly predict that variations of neighborhood density should not affect the recognition speed for the highest frequency word of any neighborhood. This prediction was tested in Experiment 1 with an auditory lexical decision task and in Experiment 2 with an auditory word naming task.

Experiment 1

Method

Subjects. Forty-three students enrolled in introductory psychology courses at Indiana University served as subjects. Subjects received course credit for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimuli. Eighty-eight words were selected for use from a computerized lexical database based on Webster's pocket dictionary (1967). From an original list of all words in Webster's pocket dictionary of seven phonemes or fewer, the final 88 words selected for the experiment were those which met the following constraints: (1) All words were listed in the Kučera and Francis (1967) corpus. (2) All words were the highest frequency words of their respective neighborhoods. Neighbors in this experiment were defined as any English word that can be derived from a target word by adding, subtracting, or deleting one phoneme (see Greenberg & Jenkins, 1964; Luce, 1986a). (3) All words had a rated familiarity of 6.0 or above on a seven-point familiarity scale, obtained from an earlier study by Nusbaum, Pisoni, and Davis (1984). In this study, all words from Webster's pocket dictionary were presented visually to subjects for familiarity ratings. The rating scale ranged from (1) "don't know the word" to (4) "recognize the word but don't know its meaning" to (7) "know the word and its meaning." The rating criterion of 6.0 and above was used to ensure that all stimulus words would be known by the subjects.

In addition to the constraints imposed for the selection of individual words, additional constraints were imposed for the separation of items into the dense and sparse neighborhood conditions. The final 44 words selected for each experimental condition were those that conformed to the following constraints: (1) The lists of words from sparse and dense neighborhoods had an equal number of words. (2) The mean frequency of all words was approximately equal across both lists. (3) The mean number of segments in all words was approximately equal across both lists.

Once the stimulus words had been selected from the set of all possible words in Webster's dictionary, the mean neighborhood density for items from the sparse category was 7 neighbors per word, and the mean neighborhood density for items from the dense category was 20 neighbors per word. The mean Kučera - Francis frequency for items from the sparse category was 2218.82, and the mean frequency for items from the dense category was 2905.14. These words are shown in the Appendix. Once all 88 words had been selected for the "word" conditions of the experiment, 88 pronounceable nonwords were generated as well. All the nonwords consisted of legal CVC monosyllables.

The stimuli were recorded in a sound-attenuated booth by a male talker of a midwestern dialect[2] using an Ampex AG500 tape deck and an Electro-Voice D054 microphone. All words were spoken in isolation. The stimuli were then low-pass filtered at 4.8 kHz and digitized at a sampling rate of 10 kHz using a 12-bit analog-to-digital converter. All words were excised from the list using a digitially-controlled speech waveform editor (WAVES) on a PDP 11/34 computer (Luce and Carrell, 1981). The mean duration of the words from sparse neighborhoods was 451.30 msec, and the mean durations for the words from dense neighborhoods was 436.80 msec. Finally, all words were stored digitally as stimulus files on computer disk for presentation to subjects during the experiment.

To ensure that all stimuli could be identified accurately, 9 additional subjects were asked to identify all the words in a perceptual identification task. All stimulus token used in the experiment were correctly identified by at least 8 of the 9 subjects in the screening test.

Procedure. Subjects were run in groups of six or fewer in a sound-attenuated room used for speech perception experiments. Each subject was seated in an individual booth equipped with a pair of matched and calibrated TDH-39 headphones and a two-button response box connected to a PDP 11/34 computer. Over each button on the response box, either a WORD or NONWORD label was situated. For half of the groups of subjects, the WORD response corresponded to the left-hand side of the response box, and for the remaining groups of subjects, the WORD response corresponded to the right-hand side of the response box. In addition to the two response buttons, a cue light was situated at the top of the response boxes to alert subjects when a trial was beginning. Subjects were instructed that they would hear brief English words or brief nonwords (examples provided were *CAT* versus *GOIP*). They were instructed to listen to each stimulus carefully and indicate whether the item was a word or nonword by pressing the appropriate button. The instructions to subjects stressed both speed and accuracy of responding.

Each trial of the experiment began with the illumination of the cue light at the top of the response box. The cuelight remained on for one second to indicate that a stimulus item was about to be presented over the headphones. Five hundred msec after the offset of the cuelight, a randomly selected spoken word or nonword was presented and the computer waited for all subjects to respond. Reaction times for each subject were recorded from the onset of the spoken stimulus until the response was executed. After all subjects responded, a 500 msec inter-trial interval elapsed, and then a new trial began. If 4000 msec elapsed on any given trial before all responses were collected, the computer recorded incorrect responses for the remaining subjects and a new trial would begin. The 176 experimental trials were preceded by 20 practice trials that were not included in the final data analysis. The practice list contained words that were not drawn from either experimental condition.

---

[2] Experiment 1 has been replicated using stimulus materials recorded by another male talker and also with stimulus materials recorded by a female talker. Patterns of results obtained in both replications closely resembled the results reported here.

Results

Mean latencies of correct responses were calculated for each subject and for each item. Reaction times shorter than 200 msec or longer than 1500 msec were excluded from calculations of means. The mean latencies for correct responses, standard deviations of mean latencies, and mean error rates for all "word" trials are shown in Table 1:

Insert Table 1 about here

As Table 1 shows, the mean latency to correctly respond to words from sparse neighborhoods was 38.07 msec faster than the mean latency to respond to words from dense neighborhoods. This difference was statistically significant by tests performed on both subjects [$F(1,42) = 64.82$, $p < .01$] and items [$t(86) = 2.36$, $p < .03$].[3]

Subset analyses:

Although the stimuli selected for Experiment 1 were carefully matched for mean word frequency and stimulus durations, examination of the items in the Appendix shows that the words from sparse and dense categories were not precisely matched for aspects of phonetic or syllabic structure. Specifically, more bisyllabic words were included in the sparse category than in the dense category. This disparity constitutes a potential confound in the results described above. Another lexical variable that had not been considered in the selection of

---

[3] There were no significant differences observed in percentages of errors between words from dense and sparse neighborhoods in any of the analyses reported in this article. Therefore, for the sake of brevity, I do not discuss error rates in any results sections.

Table 1

Mean lexical decision latencies (RT), standard deviations of latencies (SD), and percentage of errors (PE) for all word stimuli in Experiment 1.

Stimulus Type	**RT**	**SD**	**PE**
Dense	796.09	73.74	3.64
Sparse	758.02	69.58	2.88
Difference in RT	38.07 ms.		

stimulus items was the mean location of the items' *isolation* or *uniqueness points* (Marslen-Wilson & Welsh, 1978; Luce, 1986b). Because bisyllabic words tend to have earlier isolation points than monosyllabic words, it is not clear whether the differences reported above arise because of the manipulation of neighborhood density only, or if the unequal isolation points may be responsible for the differences. To address this problem, subsets of consonant-vowel-consonant (CVC) words were selected from the categories of words from dense and sparse neighborhoods. All of the 18 CVC's from the sparse category were selected, and reaction times for those items were compared to reaction times for 18 CVC's selected from the dense category. The CVC items for the dense category subset were selected to match the CVC items from the sparse category subset as closely as possible on word frequency and stimulus duration. The mean frequency of words in the sparse and dense subsets were 1091.42 and 1467.89, and the mean duration of words from the sparse and dense subsets were 528.20 and 524.80 msec, respectively.

The mean latencies for correct responses, standard deviations of mean latencies and mean error rates for all "word" trials for the subsets of CVC stimuli are shown in Table 2:

Insert Table 2 about here

Distributions of reaction times to words from dense and sparse neighborhoods were calculated for all subjects and all items. For this subset of the original data, the mean latency to correctly respond to words from sparse neighborhoods was 49.16 msec faster than the mean latency to respond to words from dense neighborhoods. The difference was significant by both tests performed on the subject means [$F(1,42) = 75.85$, $p < .01$] and the item means [$t(34) = 2.11$, $p < .05$].

The results of Experiment 1 suggest that neighborhood density can indeed affect the speed of spoken word recognition, even if the words in question are the highest frequency members of their respective lexical neighborhoods. This finding contradicts the prediction of search-based models, which posit that a self-terminating search should encounter the highest frequency word immediately and then stop processing. The results of any single experiment should always be considered with a measure of caution, however. This may be especially true for word recognition experiments, because there is currently considerable debate regarding the generality of certain experimental paradigms. Experiment 1 employed the lexical decision paradigm, which has been the topic of recent controversy in the literature (see Balota & Chumbley, 1984; Paap, McDonald, Schvaneveldt, & Noel, 1986; or Andrews, 1989, for discussion). In order to assess the generality of the findings of Experiment 1, a second experiment was conducted. This study employed a word naming paradigm instead of the lexical decision paradigm.

Table 2

Mean lexical decision latencies (RT), standard deviations of latencies (SD), and percentage of errors (PE) for a subset of the word stimuli in Experiment 1.

Stimulus Type	RT	SD	PE
Dense	831.07	68.79	3.02
Sparse	781.91	64.21	2.71
Difference in RT	49.16 ms.		

Experiment 2

Method

Subjects. Forty-one students enrolled in introductory psychology courses at Indiana University served as subjects. Subjects received course credit for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimuli. The stimuli for Experiment 2 consisted of the real-word stimuli used in Experiment 1. However, because reaction times collected in the naming task are influenced by the phonetic composition of words, several items were discarded from both the sparse and dense categories in order to balance the categories for word-initial phonemes. After the lists were equated for initial phonemes, 72 words remained, 36 from each category. As in Experiment 1, the words assigned to each category were matched as closely as possible on mean word frequency and mean stimulus duration. The average word frequency for items from the sparse category was 1322.21, whereas the average word frequency for items from the dense category was 1484.90. The average stimulus duration for items from the sparse category was 516.35 msec, and the average duration for items from the dense category was 509.22 msec.

Procedure. Subjects were tested individually in a sound-attenuated room used for speech perception experiments. Stimuli were presented over matched and calibrated TDH-39 headphones at 75 dB (SPL). A PDP 11/34 computer was used to present the stimuli and to control the experimental procedure in real-time. The digitized stimuli were reproduced using a 12-bit digital-to-analog converter and were low-passed filtered at 4.8 kHz.

All subjects were tested under the same conditions. Subjects were instructed that they would hear individual English words spoken over the headphones and that their task was to accurately repeat back each word as quickly as possible. Response latencies were collected by an Electro-Voice D054 microphone attached to a timing device accurate to within one millisecond. The timing device was connected to the PDP 11/34 computer, which collected and stored the response latencies for each subject for later analysis. Errors were recorded by the experimenter, who indicated to the computer whether or not the word presented on each trial was repeated accurately. Subjects received ten practice trials, during which the experimenter encouraged them to respond faster, if possible. The experiment proper consisted of 72 trials. The order of presentation of words within each list varied randomly across sessions.

Results

Subjects' responses were scored as correct only if the word spoken by the subject matched the stimulus word exactly. Mean latencies of correct responses were calculated for each subject and for each item. As in Experiment 1, any reaction times shorter than 200 msec or longer than 1500 msec were excluded from calculation of means. The mean latencies for correct responses, standard deviations of mean latencies, and mean error rates for all trials are shown in Table 3:

Insert Table 3 about here

As Table 3 shows, the mean latency to name words from sparse neighborhoods was 26.07 msec faster than the mean latency to name words from dense neighborhoods. This difference was significant by tests performed on both subject [$F(1,40) = 17.31$, $p < .01$] and item means [$t(70) = 2.09$, $p < .05$].

Subset analyses:

As discussed in the description of results from Experiment 1, an analysis of a subset of consonant-vowel-consonant words from the sparse and dense categories was conducted. This subset analysis was conducted in order to avoid the potential confound of differential isolation points across categories. Because several items had been removed from the total set of stimulus items, the subsets selected for comparison in Experiment 2 consisted of 15 items each. As in Experiment 1, the CVC items from both the sparse and dense categories were selected to match each other as closely as possible on word frequency and stimulus durations. The mean frequency of items from the sparse and dense subsets were 1155.80 and 1331.59, respectively, and the mean durations of items from the sparse and dense subsets were 504.33 and 508.10 msec, respectively.

The mean latencies for correct responses, standard deviations of mean latencies and mean error rates for the subset stimuli are shown in Table 4:

Insert Table 4 about here

For this subset of the original data, the mean latency to correctly respond to words from sparse neighborhoods was 32.44 msec faster than the mean latency to respond to words from

Table 3

Mean word naming latencies (RT), standard deviations of latencies (SD), and percentage of errors (PE) for all stimuli in Experiment 2.

Stimulus Type	RT	SD	PE
Dense	570.72	36.14	0.47
Sparse	544.65	60.04	0.66
Difference in RT	26.07 ms.		

Table 4

Mean word naming latencies (RT), standard deviations of latencies (SD), and percentage of errors (PE) for a subset of the stimuli in Experiment 2.

Stimulus Type	**RT**	**SD**	**PE**
Dense	583.93	25.64	0.56
Sparse	551.49	57.45	0.22
Difference in RT	32.44 ms.		

dense neighborhoods. This difference was significant by the test performed on the subject means [$F(1,40) = 14.97$, $p < .01$], but the difference only approached significance by the test of the item means [$t(28) = 2.04$, $p < .06$].

General Discussion

The present experiments were conducted to test the prediction derived from search-based models of spoken word recognition that neighborhood density effects should not be observed for the highest frequency word of any given neighborhood. Contrary to this prediction, reliable effects of neighborhood density were observed for words from sparse and dense neighborhoods in both lexical decision and naming tasks. By testing the strong predictions generated by search models, we may evaluate the class of models against more general activation-based models. Given the results obtained in the present experiments, as well as those reported by Luce (1986a) and by Andrews (1989), it is appropriate to survey several general classes of word recognition models and consider how adequately they fare with respect to neighborhood density effects. Three kinds of models are considered in turn: serial search models, parallel unlimited-capacity models, and parallel limited-capacity models.

Serial Search Models:

Given the findings reported by Andrews (1989) for visual word recognition, and the results of the present experiments for spoken word recognition, it is apparent that frequency-ordered search-based models are too rigid to accurately predict the effects of neighborhood density on word recognition. For printed words, whereas search models such as Forster's (1976) model or the activation-verification model (Becker, 1976; Paap et al., 1982; 1986) predict that low frequency words should be *inhibited* by large neighborhoods, such words are actually *facilitated* by their neighborhoods. For spoken words, whereas search models predict that there should be *no effects* of neighborhood density on the highest frequency word of any neighborhood, such words are actually *inhibited* by their neighborhoods.

Neighborhood density effects on word recognition present a dilemma for serial search models, and it is difficult to imagine how these findings may be reconciled. For example, we may consider several options for modifying Forster's (1976) search model to account for the present findings: First of all, we could assume that the search process is actually better described as exhaustive (or partially exhaustive) rather than self-terminating, and that the duration of search is mediated by set size. While this approach would help resolve the model's disparity with the present results, it does not seem representative of the "true spirit" of the model. With a loosening of the original self-terminating search assumption, the model's best characteristics- its testability and strong position regarding frequency effects- would be sacrificed. In addition, it would be far more difficult to predict frequency effects without the self-terminating search assumption. Another possible approach that may work better is to modify the model's matching and decision procedures. For instance, one could assume that the model's best-match comparisons are affected by each word's overall similarity to its nearest neighbors, and that these similarity relations may be more influential and delay search more for words from dense neighborhoods than for words from sparse neighborhoods. Unfortunately, until more is known about the precise similarity relations that exist among words from various neighborhoods with different kinds of phonetic constitutions, there is little empirical evidence to justify this modification of search models.

Parallel, Unlimited-capacity Models:

Under the general heading of activation-based models, the first class of models to consider are *parallel, unlimited-capacity* models. A familiar example is Morton's (1969; 1979) logogen model. The logogen model is a simple parallel-processing system that integrates multiple sources of information to determine the candidacy status of any number of words. Logogens receive bottom-up excitation from sensory information as well as top-down excitation from contextual information. Word frequency is treated in the model as gradations in the resting activation levels of the logogens for words of differing frequencies. In all these respects, the logogen model closely resembles the interactive activation model (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). However, unlike the interactive activation framework, which posits lateral inhibitive connections between word nodes, Morton's model assumes independence among all logogens. The independence among logogens in Morton's model implies that an *unlimited* processing capacity is available for word recognition. Clearly, this unlimited capacity assumption is inconsistent with findings of neighborhood density effects since the logogen model assumes that no influence is exerted among lexical neighbors (see Luce, 1986a).

Another model that embodies the assumption of unlimited-capacity parallelism is Marslen-Wilson's cohort theory (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), although cohort theory is better able to reconcile itself with neighborhood effects than the logogen model. In a recent paper, Marslen-Wilson (1987) states that cohort theory assumes that any number of word candidates may be activated in parallel with no consequences of set size for the processor. For some words, cohort model avoids the problems inherent in this unlimited-capacity assumption by asserting the fundamental role of variable word recogntion points. By acknowledging the role of recognition points, Marslen-Wilson's model captures the importance inherent in the structural relations among sound patterns in the lexicon (Pisoni & Luce, 1987). Unfortunately, despite the postulation of recognition points, cohort theory's unlimited-capacity assumption still has problems. For instance, among short words, such as the CVC words used in the present experiments, recognition points simply do not differ enough to account for the neighborhood density effects that have been repeatedly observed (see also Luce, 1986a; 1986b; Luce, Pisoni, & Goldinger, in press). Therefore, it appears that the unlimited processing capacity proposed in both logogen theory and cohort theory may be an unwarranted assumption.

Parallel, Limited-capacity Models:

The second class of models to consider within the activation-based class of models are *parallel, limited-capacity* models. The example that has been discussed throughout this paper is the interactive activation model (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). As a rough approximation, the interactive activation model is similar in principle to Morton's (1969) logogen model, but there are two major differences that distinguish the models: First of all, the interactive activation model is simply much more explicit than the logogen model with regard to actual processing assumptions. Second, the interactive activation model posits lateral inhibitory connections between nodes resident at the same level of the architecture, such as between word nodes. By assuming lateral inhibition, the interactive activation model is functionally limited in its processing capacity; the overall excitation level among word nodes has a damping effect on all individual items. By recourse to this feedback mechanism, the interactive activation model can account for the effects of neighborhood density reported by Luce (1986a), Andrews (1989), and in the present experiments.

A second model that falls into the parallel, limited-capacity category is the neighborhood activation model, described by Luce (1986a; Luce, Pisoni, & Goldinger, in press; Goldinger, Luce, & Pisoni, 1989). The neighborhood activation model posits that, upon stimulus input, a neighborhood of word candidates in the lexicon is activated. As the neighborhood's candidates are activated, *word decision units* dedicated to each lexical entry conduct best-match comparisons to resolve which of the candidates was actually presented. The word decision units are biased by word frequency (as opposed to integrating frequency into initial activation levels, as in logogen theory), and are mutually influential. In the neighborhood activation model, each decision unit is assumed to be sensitive to the overall activity in the decision system. Therefore, like the interactive activation model, the neighborhood activation model is functionally limited in its processing capacity and is able to account for effects of neighborhood density on word recognition.

Given this brief survey of models of word recognition, it appears that the parallel, limited-capacity models are most adequate for explaining the range of phenomena previously discussed in the literature, as well as the effects of neighborhood structures that have been discussed more recently. However, the explanatory adequacy of models such as the interactive activation model comes at the high cost of testability (Andrews notes that the interactive activation model "wins by default" (1989, pg. 811)). Future research should be dedicated to testing the interactive models more explicitly. For the present time, there are still numerous issues regarding neighborhood effects that require further investigation. First of all, in the visual word recognition literature, Grainger, O'Regan, Jacobs, & Segui (1989) have argued that neighborhood *frequency* may be the primary determinant of recognition time; rather than neighborhood density. Similarly, Luce (1986a) reported finding large and significant effects of neighborhood frequency on spoken word recognition. Further work should address the relative importance of these neighborhood characteristics. Another question that needs to be more rigorously investigated is the underlying nature of neighborhood density effects, as mentioned above. It is not entirely clear from the data reported in the present experiments whether what we refer to as "neighborhood density" effects are best explained by the actual *number* of words in a neighborhood, or by the given target words' overall similarity or *closeness* to its neighbors, or perhaps its nearest neighbor. Hopefully, as more exact methods are developed to quantify the degrees of perceptual similarity among words, these questions may be answered more definitively.

Finally, even within the powerful framework of the interactive activation model, it is not immediately apparent why orthographic neighborhoods should cause facilitation for visual word recognition whereas acoustic-phonetic neighborhoods should cause inhibition for auditory word recognition. Presumably, the interactive activation model could simulate these disparate effects by differentially tuning the excitatory and inhibitory parameters for nodes that correspond to orthographic and phonological feature input. Regardless of the model's ability to mimic the effects of both visual and auditory neighborhoods, however, it is clearly more important for the interactive activation framework to provide an adequate theoretical account for these modality differences. The problem may be even more difficult for totally distributed models of word recognition, such as the recent Seidenberg and McClelland (1989) model.

In summary, the present experiments have shown that neighborhood density affects the speed of spoken word recogntion, even though the words employed were the highest frequency words of their respective neighborhoods. This finding is inconsistent with predictions derived from search-based models of word recognition which assume that no differential effects of neighborhood density should be observed for such privileged words. Furthermore, the present findings provide a spoken-word analog to Andrews' recent (1989) investigation of visual word recognition, and they corroborate her conclusions regarding the superiority of activation-based models over serial models of word recognition.

References

Adams, M.J. (1979). Models of word recognition. *Cognitive Psychology*, **11**, 133-176.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 802-814.

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Balota, D.A., & Chumbley, J.I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, **24**, 89-106.

Becker, C.A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 556-566.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI*, (pp. 535-555). Hillsdale, NJ: Erlbaum.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales & E. Walker (Eds.) *New approaches to language mechanisms*. Amsterdam: North Holland.

Forster, K.I., & Bednall, E.S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, **4**, 53-61.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Grainger, I.J., O'Regan, J.K., Jacobs, A.M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics*, **45**, 189-195.

Greenberg, J.H., & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, **20**, 157-177.

Kučera, F., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Luce, P.A. (1986a). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A. (1986b). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, **39**, 155-158.

Luce, P.A., & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on speech perception progress report no. 7.* Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (in press). Similarity neighborhoods of spoken words. In Altmann, G. (Ed.), *Cognitive representation of speech*, Cambridge: MIT Press.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Mason, M. (1975). Reading ability and letter search time: Effects of orthographic structure defined by single-letter positional frequency. *Journal of Experiment Psychology: General*, **104**, 146-166.

Massaro, D.W., Venezky, R.L., & Taylor, G.A. (1979). Orthographic regularity, positional frequency, and visual processing of letter strings. *Journal of Experimental Psychology: General*, **108**, 107-124.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Morton, J. (1979). Word recognition. In Morton, J., & Marshall, J.C. (Eds.), *Structures and processes*, (pp. 109-156), Cambridge: MIT Press.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Paap, K.R., McDonald, J.E., Schvaneveldt, R.W., & Noel, R.W. (1986). Frequency and pronounceability in visually presented naming and lexical decision tasks. In Coltheart, M. (Ed.), *Attention and performance XII: The psychology of reading*, 221-243.

Paap, K.R., Newsome, S.L., McDonald, J.E., & Schvaneveldt, R.W. (1982). An activation-verification model for letter and word recognition: The word-superiority effect. *Psychological Review*, **89**, 573-594.

Pisoni, D.B., & Luce, P.A. (1987). Acoustic-phonetic representations in spoken word recognition. *Cognition*, **25**, 21-52.

Rumelhart, D.E., & McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60-94.

Seashore, R.H., & Eckerson, L.D. (1940). The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, **31**, 14-38.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523-568.

Tyler, L.K., & Fraunfelder, U.H. (1987). The process of spoken word recognition: An introduction. *Cognition*, **25**, 1-20.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Appendix: Word stimuli used in Experiment 1.

Sparse neighborhoods:

earth, of, up, were, able, allow, and, any, ask, blew, chief, church, death, dog, early, easy, echo, else, evil, five, food, give, honor, idle, iron, item, judge, love, move, occur, okay, old, other, over, power, sky, south, teeth, these, this, thought, voice, was, young

Dense neighborhoods:

age, are, each, he, in, know, to, back, both, but, did, does, door, down, face, firm, for, full, get, girl, had, job, keep, less, like, long, one, peace, pool, put, real, right, road, rock, said, serve, shall, shape, some, that, their, top, with, work

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Contrast and Normalization in Vowel Perception[1]

Keith Johnson

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This paper will appear in *Journal of Phonetics*. The comments and criticisms of John Mullennix, Van Summers, David Pisoni, Neal Johnson, Anthony Bladon and an anonymous reviewer are gratefully acknowledged. Ying Yong Qi wrote an earlier version of code to implement the Patterson filters which I adopted here. The research was supported by NIH Training Grant No. NS-07134-11.

Abstract

The work reported in this paper is an attempt to better understand vowel normalization by investigating the relationship between vowel normalization and vowel contrast. In the first experiment, vowels from a "hood"-"hud" continuum were presented at two levels of fundamental frequency (F0). In one condition, tokens were blocked by F0, in the other, tokens with different F0 levels were randomly intermixed with each other (as in the typical F0 normalization experiment). Subjects identified the tokens in quite different ways depending upon the type of presentation. In the mixed presentation, they identified the high F0 items most often as "hood" and the low F0 items most often as "hud". In the blocked condition, there was no reliable difference between the high and low F0 continua. This pattern of results suggests that a contrast effect is at work. The second section of the paper reports a series of simulations in which four models of perceptual contrast are tested. Auditorily-based (AB) spectra served as inputs to the simulations. The AB spectra were produced by a model which incorporates two levels of processing, (1) narrow-band auditory filtering (Patterson, 1976) and (2) wide-band integration (Chistovich, 1985). Results of the first experiment could be approximated by two models: an auditory figure/ground model, and a talker contrast model. A second experiment tested these two models. The auditory figure/ground model predicts that in a cross-series anchoring experiment (in which tokens with high F0 are used to anchor the low F0 continuum and tokens with low F0 are used to anchor the high F0 continuum) the boundary of the vowel identification function will be shifted toward the vowel quality of the anchoring stimulus. The talker contrast model predicts that the vowel quality of the anchoring stimulus is less important than its F0 and that the phoneme boundary will be shifted in the same direction regardless of the vowel quality of the anchoring stimulus. The results of the experiment quite unambiguously supported the predictions of the talker contrast model.

Contrast and Normalization in Vowel Perception

Typically vowel normalization and vowel contrast have been studied separately. Researchers interested in vowel contrast have generally not considered vowel normalization processes (although see Fox, 1985), and researchers who have investigated vowel normalization have not considered the possible role of perceptual contrast in the process of vowel normalization. This is especially interesting in view of the fact that fundamental frequency (F0) normalization experiments involve intermixing tokens with different F0. The research reported here is an attempt to gain a better understanding of vowel normalization (and, to some extent, vowel contrast) by determining the ways in which perceptual contrast and vowel normalization interact. The research is motivated by the assumption that we may be able to come to a better understanding of both of these important perceptual processes by considering the ways in which they interact.

Vowel Normalization

Vowel normalization is a hypothetical perceptual process in which interspeaker vowel variability is reduced in order that perceptual vowel identification may then be performed by reference to relative vowel quality rather than the absolute values of the acoustic parameters of vowels. It is well documented that hearers are influenced by F0 when they make judgements about vowel quality (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968; Ainsworth, 1975; Traunmüller, 1981). The general pattern of results reported in these studies is that the vowel formants must increase as F0 is increased in order to maintain the same vowel quality. Because there is a correlation between vocal tract size and F0, it is also possible to describe the effect in terms of the perception of the size of the talker's vocal tract. In this case, we would say that hearers perceptually normalize vowel formants by reference to some index of vocal tract size, using F0 as a cue for this variable. Many researchers have assumed a similar process of normalization (the utilization of "an internal model of the speaker", Summerfield, 1971) and so, for this reason, the effect has been called vocal tract normalization.

As a cue for perceived speaker identity, F0, in this view, may be used by hearers to establish a speaker-dependent perceptual vowel space. In this way, the expected acoustic correlates of particular vowel qualities are adjusted on the basis of the perceived identity of the talker. I will call this type of hypothesis an *adjustment-to-talker* model of vowel normalization. One way to implement an adjustment-to-talker model of vowel normalization is to use F0 in order to estimate speaker-dependent formant ranges in a range normalization process (concerning range normalization see Gerstman, 1968). The model of vowel normalization described by Bladon, Henton and Pickering (1984) is also an example of an adjustment-to-talker approach. This is not obvious, at first, because they describe the model as "auditory", but it is actually a two-stage model. They propose that vowel normalization is accomplished by shifting the auditory spectra of vowels produced by female speakers down by 1 Bark before comparing them to spectral templates based on vowels produced by men. The auditory stage in the model is in the calculation of "auditory" spectra (Bladon and Lindblom, 1981). The second stage (when spectra are shifted along the Bark scale) involves an adjustment-to-talker.

Traunmüller (1981) proposed an explanation of F0 normalization which is an alternative to the adjustment-to-talker view of normalization in that it appeals only to properties of the auditory system. No information about speaker identity or vocal tract length need be specified in this model.[1] He proposed that the 'centre of gravity' effect reported by Chistovich, Sheikin and Lublinskaja (1979) can account for vowel normalization data.

Chistovich et al. (1979) reported that vowel formants which are within about 3 Bark of each other seem to be integrated into a single perceptual 'centre of gravity'. In their experiments, subjects manipulated the frequency of a synthetic single-formant vowel until it matched, as closely as possible, the quality of a two-formant standard. They found that when the formants of the standard were within about 3 Bark of each other, subjects tended to adjust the frequency of the single-formant so that it fell at the amplitude weighted mean of the formants of the standard. However, when the separation between the formants of the standard was greater than 3 Bark, subjects tended to adjust the single formant so that it matched one of the formants of the standard. These data are consistent with the observations of Delattre, Liberman, Cooper and Gerstman (1952) in which it was reported that acceptable back vowels, in which F1 and F2 are close in frequency, can be synthesized with only a single formant. However, for front vowels, in which F2 and F3 are near each other, the F2 of two-formant synthetic vowels must be higher than the F2 found in natural speech. The data reported in Johnson (1989) are also consistent with the 'centre of gravity' effect. Johnson reported that a higher formant normalization effect (see Fujisaki and Kawashima, 1968) is found when F2 and F3 are within 3 Bark of each other, but not when F2 and F3 are separated by more than 3 Bark.[2]

Traunmüller's proposal is that the effect of increasing F0 on vowel perception results from an increased participation of the lowest harmonic in the 'centre of gravity' which corresponds to the perceived F1 (F1'). As F0 increases the lowest harmonic enters more and more into the window of integration which includes the peak of F1. This increased influence of the lowest harmonic results in a lower F1' as F0 increases. Of course, as F0 increases past the actual F1, F1' follows F0. In the second section of this paper, we will consider whether spectra generated by an auditory model have the properties suggested by Traunmüller.

---

[1] Other approaches to perceptual vowel normalization which avoid reference to the size of the vocal tract include those discussed by Sussman (1986), Syrdal and Gopal (1986), Nearey (1978) and Miller (1989). This research and that of Gerstman (1968) and Labonov (1971) (see Disner, 1980 for a review and critique) is concerned with normalization algorithms, and not explicitly with perceptual processing. I have chosen to emphasize models of vowel normalization which are more directly concerned with perceptual processing, but this does not indicate a disregard for the algorithm approach. Rather, I wish to avoid some of the assumptions of this approach: in particular the assumption that vowel normalization is preceeded by formant extraction. Clearly, vowel normalization algorithms which correctly classify vowels may reflect perceptual processes which operate in ways similar to those encoded in the algorithms (see Traunmüller, 1981; Nearey, 1978; and Sussman, 1986).

[2] Although the criticisms of the use of front vowel continua (see the discussion of Experiment 2) also apply to the work in Johnson, 1989.

Vowel Contrast

One of the earliest findings in the study of speech perception was that vowel discrimination is better than would be predicted by vowel labelling (Fry, Abramson, Eimas and Liberman, 1962, Eimas, 1963). These authors also found that the label given to an ambiguous vowel token is a function of context. When an ambiguous token from a continuum from A to B is preceeded by the A endpoint the ambiguous token is more likely to be labeled B, and in the context of the B endpoint the label will be shifted to A. This *vowel contrast* effect, coupled with better-than-predicted discrimination led the Haskins group to conclude that vowels are perceived continuously because small changes in vowel quality may be easily produced, while consonants are perceived more categorically because their production tends to be categorical. The work of Fujisaki and Kawashima (1969) and Pisoni (1971, 1973, 1975) turned attention to the role of auditory memory in speech perception. It was hypothesized that vowels, which have longer, more steady-state acoustic cues, leave a longer lasting trace in auditory memory, and so the auditory differences between vowels are more readily available for use in speech perception tasks. Thus, a dual-process view of vowel perception (which involves both an auditory stage and a phonetic stage) provides an explanation of vowel context effects. Simon and Studdert-Kennedy (1978) called this view of vowel contrast an auditory figure/ground approach.

Feature detector theory (Eimas, Cooper and Corbit, 1973; Eimas and Corbit, 1973; Cooper, 1974) provides a possible explanation of vowel contrast in terms of feature-detector fatigue. "However, this hypothesis has rarely been mentioned in connection with vowel perception, presumably because the large number of vowel categories and the relatively noncategorical perception of the stimuli made explanations in terms of discrete dectors seem unattractive. Also, while feature-detector fatigue is a plausible mechanism for explaining selective adaptation effects, it cannot account for pairwise contrast where only a single contextual item is presented" (Fox, 1985, p. 1552).

Crowder (1981) proposed a model of vowel contrast in which the memory representations behave "in accordance with the laws of *recurrent lateral inhibition*" (p. 175). In this model, the auditory memory representation of a stimulus interacts with that of an earlier stimulus in a process which is analogous to lateral inhibition in peripheral sensory systems. Crowder suggests that the frequency components of memorial representations of stimuli interact with each other in auditory memory in such a way that unique components are relatively uninhibited and tend to dominate in the classification of stimuli, and overlapping frequency components are mutually inhibiting. Thus, in Crowder's view of vowel contrast, the spectral differences between stimuli are enhanced and similarities are inhibited as a result of the nature of their representations in auditory memory and the hypothesized process of lateral inhibition. An important feature of Crowder's proposal is the notion of "channels" in auditory memory. According to Crowder, items produced by different talkers will occupy different channels of auditory memory, with degree of perceived talker difference determining the degree of channel discrepancy. He suggests that items which are on the same or similar channels will inhibit each other, while items on different channels will not. Therefore, it is not clear how relevant this model of vowel contrast is to situations in which vowels with different speaker qualities are presented.

Simon and Studdert-Kennedy (1978) and Fox (1985) also considered a response bias explanation (Parducci, 1965, 1975) of vowel contrast. In this account, the change in labelling behavior in an anchoring experiment is due to the increase in the number of stimuli which must receive one label. Parducci's range-frequency theory predicts that subjects will attempt to use category labels an equal number of times during an experimental session, and so will show a boundary shift when anchoring stimuli are presented with a test continuum, because the anchor stimuli are consistently labelled with one of the available labels. This seems to explain the results of anchoring experiments where one stimulus is presented more often than another, but it does not account for evidence of vowel contrast in experiments where each of the stimuli is presented equally often (Fry, et al., 1962). The fact that there is a shift in identification in anchoring experiments, even when subjects are made aware of the relative frequency of occurrence of each token, also suggests that the response bias explanation is not an adequate explanation of this effect (for other arguments see Fox, 1985, p. 1553).

Finally, in addition to vowel contrast effects, we will consider in this paper the possibility that the perceived identity of the speaker may be subject to talker contrast effects. That is, when two (synthetic) voices are placed close to each other in time, the degree of perceived difference between the voices may be larger than when they are temporally separate. As will be shown below, this type of contrast is an important consideration for experiments, such as those reported here, in which both vowel quality and speaker quality are manipulated.

Experiment 1

Most previous studies of F0 normalization have one methodological trait in common: stimuli at different F0 levels are presented intermixed with each other. This presentation format corresponds to one of the conditions studied by Mullennix, Pisoni and Martin (1989). In those experiments, they found that when hearers were required to identify words in different levels of noise, word recognition performance was impared by random variation of talker identity. Performance was better when all of the words presented for identification had been produced by the same talker, as compared to a condition in which the identity of the talker varied from trial to trial. Mullennix et al. also found reliable reaction time differences between single-talker and multiple-talker conditions in two naming experiments. Subjects could repeat aloud words in the single-talker condition about 50 ms faster than they could in the multiple-talker condition (averaged over lexical density and word frequency conditions in two experiments). The reaction time data were interpreted as indicating that hearers must adjust to the talker in the multiple-talker condition while this adustment is not required in the single-talker case. Mullennix et al.'s experiments are relevant to the study of vowel normalization because they indicate that hearers do not automatically "normalize" the speech that they hear, but rather that some exposure to a new talker is required in order to be able to identify words as quickly and accurately as possible.

Previous research on F0 normalization has involved the presentation of synthetic speech tokens in what is essentially a multiple-talker condition, though this "multiple-talker" condition has usually been composed of only two levels of F0 (Miller, 1953; Fujisaki and Kawashima, 1968; and Slawson, 1968). The present experiment extends the traditional format by including a condition in which tokens are blocked by F0. In the analog of Mullennix et al.'s single-talker condition (here called the single-F0 condition), the tokens were blocked by F0, thus the F0 of the tokens was entirely predictable within blocks. In the analog of their multiple-talker condition (here called the mixed-F0 condition), the tokens of the two F0 continua were randomly intermixed with each other.

Method

Subjects. Twenty-four undergraduate students at Indiana University participated in the experiment (18 female, 6 male). All were native speakers of American English who had never experienced any speech or hearing disorders. They received partial course credit in an introductory psychology course for their participation.

Materials. The stimuli used in this experiment were synthetic CVC syllables in a vowel continuum from [hɷd] to [hʌd]. Two continua were synthesized (using the Klatt, 1980 cascade-parallel formant synthesizer) - one with a steady-state F0 of 120 Hz, the other with steady-state F0 of 240 Hz. The formant values of the synthetic vowels are shown in Table 1 and in Figure 1. These formant values had been used in previous studies of vowel F0 normalization (Johnson, 1989, in press). The syllables were 285 ms in duration. The aspiration noise of the /h/ was 95 ms long (with F1 and F2 slightly higher than the F1 and F2 of the vowel as naturally occurs as a result of tracheal coupling). The steady-state vowel portion of the stimuli was 160 ms long. Bandwidths of F1-F3 were 110, 75 and 110 Hz, respectively. F4 and F5 were 3500 Hz and 4200 Hz, both with a bandwidth of 300 Hz, and were steady-state throughout the syllables. The final transitions into /d/ were 30 ms long and ended at 300, 1700 and 2516 Hz for F1-F3, respectively. The F3 transition dipped to

2116 over the first 15 ms of the transition and then rose to 2516. The peak amplitudes of the tokens were equated (to avoid possible effects of amplitude variation on reaction time).

Insert Table 1 about here

Insert Figure 1 about here

Procedure. Stimuli from the two vowel continua were presented over TDH-39 headphones at a listening level of 80 dB using two types of presentation. In the single-F0 condition, the tokens were blocked by F0. In the mixed-F0 condition, the tokens from the two continua were randomly intermixed with each other. Each stimulus was presented ten times in each of these two conditions. Subjects were randomly divided into two groups. One group of subjects responded first to the items in the single-F0 condition and then to the same items in the mixed-F0 condition. The other group first heard the items in the mixed-F0 presentation type and then responded to them again in the single-F0 condition. The two groups will be called the single-first and mixed-first groups, respectively. In the single-F0 condition, the order of presentation of F0 level was counter-balanced across subjects. So, presentation type and F0 level were treated as within-subject variables while order of presentation was treated as a between-subjects variable.

A video monitor was mounted at approximately eye level for each subject. The words HOOD and HUD were presented on the monitor at the left and right of the screen. Subjects used these labels as an indication of which button to press in a forced-choice identification task. For half of each subject's responses in each condition, HOOD was the right-hand response and HUD was the left-hand response. For the other half of the trials, the right-hand response was HUD and the left-hand response was HOOD. Button to response associations were switched at intervals of 70 trials, so within a block of 70 trials the association was constant.

Each token in the two continua was presented 10 times in each of the two types of presentation. The number of presentations per subject was 280 (7 tokens * 2 F0 levels * 2 presentation types * 10 presentations). Both identification and reaction time data were collected online by a PDP 11/34 mini-computer. Subjects participated in the experiment in groups of up to six at a time.

Table 1

Formant values of the test tokens used in the listening experiment.

Token #	1	2	3	4	5	6	7
F1	474	491	509	526	543	561	578
F2	1111	1124	1137	1150	1163	1176	1189
F3	2416	2424	2432	2440	2448	2456	2464

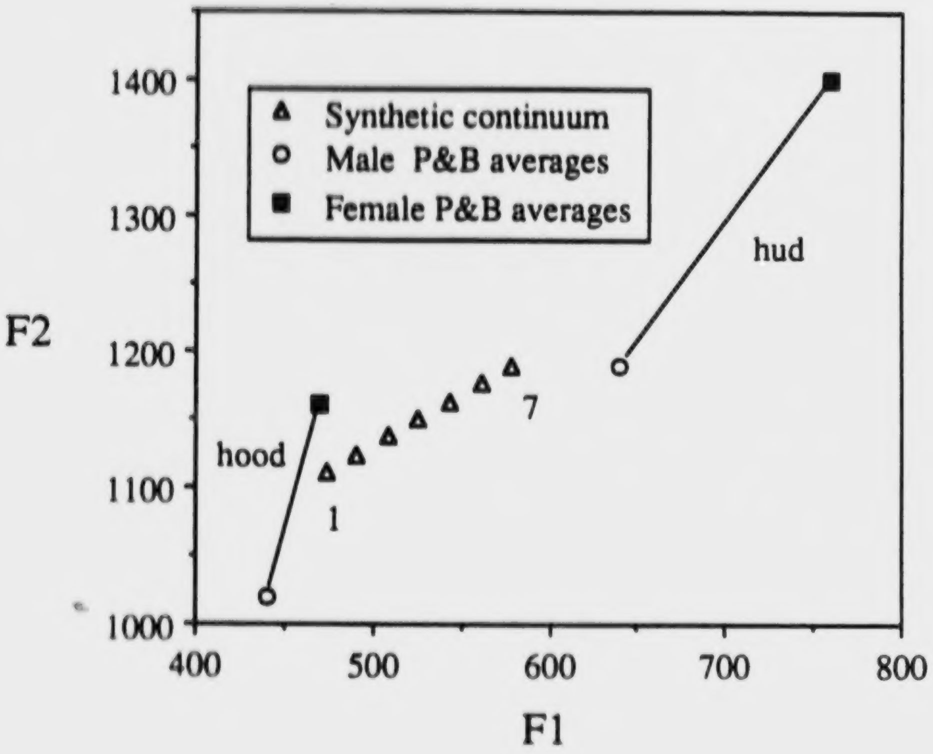


Figure 1. Tokens in the "hood"-"hud" continuum, and the Peterson and Barney (1952) average values for midwestern male and female speakers of English.

Results

Identification Data. The identification data are shown in Figure 2. The two presentation conditions are shown in separate graphs, the responses of the mixed-first group are plotted with open symbols, while the responses of the single-first group are plotted with filled symbols. Responses to the low F0 continuum are plotted with circles and the responses to the high F0 continuum are plotted with squares. The identification data were analyzed in a four-way repeated measures ANOVA with factors PRESENTATION TYPE (mixed-F0 vs. single-F0), F0 LEVEL (120 Hz vs. 240 Hz.), GROUP (mixed-first vs. single-first) and TOKEN (1-7).

Insert Figure 2 about here

The only significant main effects in the analysis were for F0 LEVEL $[F(1,22) = 91.76, p < .0001]$ and TOKEN $[F(6,132) = 154.14, p < .0001]$. The low F0 continuum was identified as "hood" 35.5% of the time while 64.9% of the high F0 items were labeled "hood". The interaction of PRESENTATION TYPE and F0 LEVEL was significant $[F(1,22) = 123.07, p < .0001]$. This interaction can be observed in Figure 2 as the difference between circles and squares in the top and bottom panels. Table 2 shows the average percent "hood" responses to each continuum (low and high F0). When items which differed in F0 were presented intermixed with each other, there was a large difference between the identification functions as a function of token F0, while when the items were presented in separate blocks there was no effect of F0 on identification behavior. The three-way interaction of PRESENTATION TYPE, F0 LEVEL and TOKEN was also significant $[F(6,132) = 4.82, p < .001]$. Examination of the functions in Figure 2 indicates that this interaction occurred because the effect of F0 in the mixed-F0 condition was to shift the boundary between "hood" and "hud" and not a global change in the probability of "hood" responses across the continuum.

Insert Table 2 about here

The interaction between PRESENTATION TYPE and TOKEN was also significant $[F(6,132) = 32.61, p < .0001]$. The average identification function in the mixed-F0 condition was flatter than was the average identification function in the single-F0 condition. Also, there was an interaction between F0 LEVEL and TOKEN $[F(6,132) = 15.33, p < .0001]$. The effect of F0 (averaged over groups and presentation conditions) was a boundary shift and not a global change in probability of "hood" response.

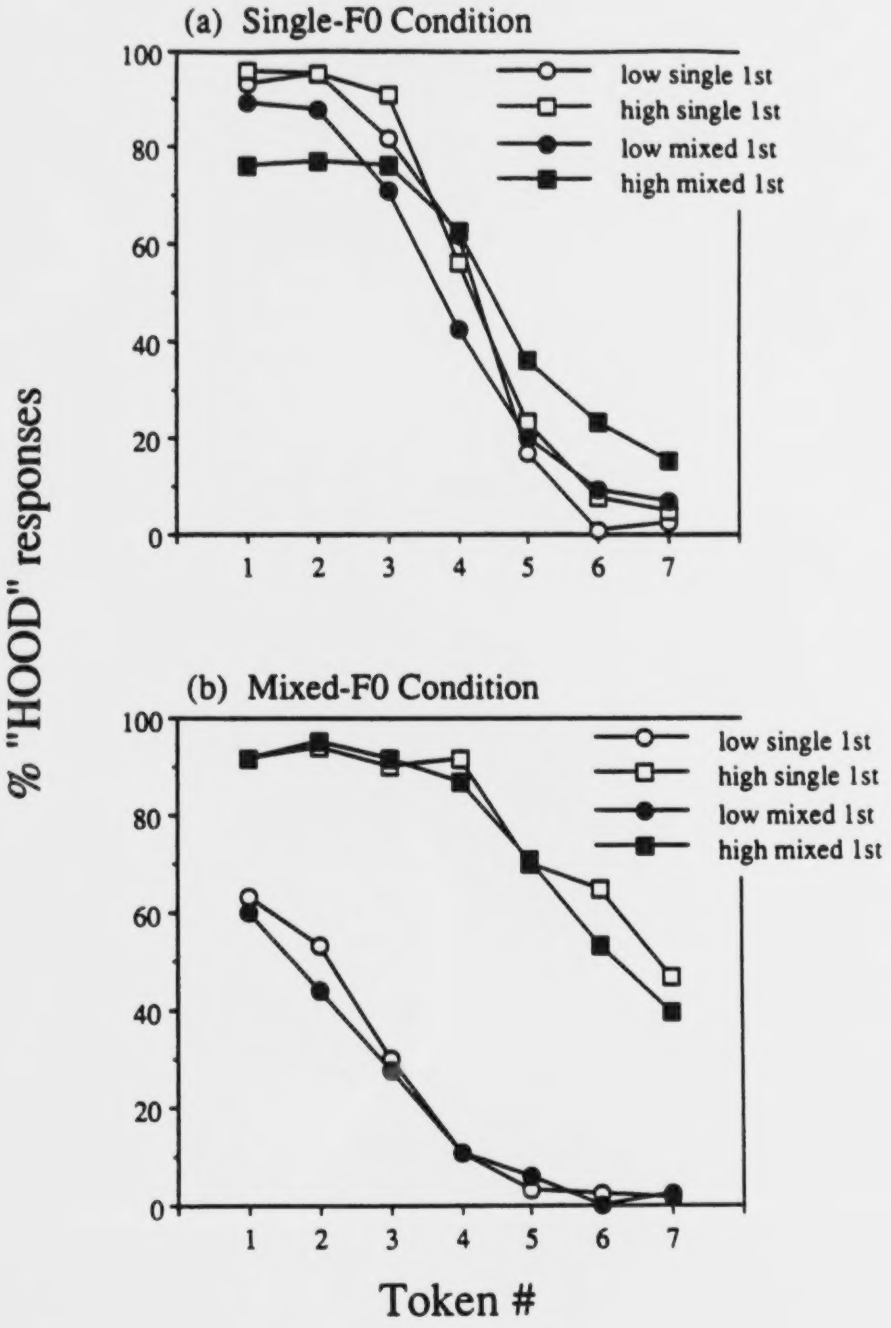


Figure 2. Identification data as a function of token number, presentation type, subject group and F0 level. The mixed-first group is plotted with open symbols. The single-first group is plotted with closed symbols. The high F0 continuum is plotted with squares and the low F0 continuum is plotted with circles.

Table 2

The interaction of PRESENTATION TYPE and F0 LEVEL. The data in this table are percent "hood" identifications as a function of presentation type and F0 level averaged across subjects and tokens.

	Low F0	High F0
Single-F0	48.4	52.7
Mixed-F0	22.6	76.9

Finally, there were two significant interactions which involved group differences. As is clear in Figure 2b, the two groups of subjects (mixed-first and single-first) had virtually identical response functions in the mixed-F0 condition, while in the single-F0 condition (Figure 2a), the response functions for the mixed-first group were somewhat flatter than those of the single-first group. This difference was reflected in the PRESENTATION TYPE by TOKEN by GROUP interaction $[F(6,132) = 4.25, p < .001]$. Also, in Figure 2a (the single-F0 condition), it appears that this group difference was larger for the high F0 level than for the low F0 level (i.e. the function for the high F0 mixed-first continuum is flatter than the function for the low F0 mixed-first continuum). This was reflected in the four way interaction of PRESENTATION TYPE, GROUP, F0 LEVEL and TOKEN $[F(6,132) = 3.02, p < .01]$. It is not clear why the subjects in the mixed-first group would show less categorical identification of the continua in the single-F0 condition than subjects in the single-first group.

Reaction Time Data. Average reaction times, measured from item onset and averaged across both response and token, were analyzed in a repeated measures ANOVA with factors F0 LEVEL, PRESENTATION TYPE and GROUP. The only statistically significant effect in this analysis was the main effect for PRESENTATION TYPE $[F(1,22) = 5.39, p < 0.05]$. Average reaction time in the mixed-F0 condition was 697 ms and in the single-F0 condition was 647 ms. The main effect for F0 LEVEL approached significance $[F(1,22) = 3.76, p = 0.0653]$. The trend was for items with low F0 to be identified more quickly than the items with high F0. This effect may relate to the relative naturalness of the different levels of F0, since the tokens sounded more natural at lower F0 levels.

An additional analysis of the reaction time data from the mixed-F0 condition assessed the effect of changing F0 from token to token. Classifying the reaction time data by the F0 of the item being identified and by the F0 of the immediately preceding item results in four classes of reaction times; low F0 items which were immediately preceded by a low F0 item, low F0 items which were immediately preceded by a high F0 item, high F0 items which were immediately preceded by a low F0 item and high F0 items which were immediately preceded by a high F0 item. Analysis of these data in a three-way repeated measures ANOVA with factors: TOKEN-F0 (high or low), CONTEXT-F0 (high or low) and GROUP (mixed-first or single-first) revealed one reliable main effect (TOKEN-F0 $[F(1,22) = 9.96, p < .01]$). This effect is consistent with the marginal effect for F0 LEVEL found in the overall analysis. There was also a significant interaction between the TOKEN-F0 and CONTEXT-F0 factors $[F(1,22) = 5.98, p < .05]$. When there was a change of F0 from one token to the next, subjects were slower to identify the token than when F0 did not change from one token to the next.

Discussion

The difference in reaction time between the mixed-F0 and single-F0 conditions which was observed in this experiment (50 ms.) is comparable to the reaction time difference observed by Mullennix et al. (1989) between their single-talker and multiple-talker conditions. A reaction time difference for blocked versus mixed voices was also reported by Summerfield and Haggard (1975). They found a reaction time difference which could be attributed to a normalization process, as opposed to a general effect of divided selectional attention, when F0 and F3 varied together, but not when F0 varied alone. Although it is possible that the reaction time difference found here reflects a normalization process (F0 variation was much greater in this study than in that of Summerfield and Haggard), the proper control conditions were not included in this study, and so it would be premature to claim that the reaction time difference is evidence for a special normalization process.

The identification data indicate that when tokens from vowel continua with different F0 are presented randomly intermixed with each other there is an effect of F0 upon vowel identification, however, when tokens are presented blocked by F0 the effect of F0 is severely diminished (if present at all). This pattern of results suggests the operation of a contrast effect. In the sections that follow we will attempt to determine what type of contrast effect can account for these data.

Model Studies of Vowel Normalization and Vowel Contrast

This section is organized into three parts. Section 3.1 is a description of a model of the auditory representation of vowels. In the section 3.2, spectral representations generated by this model are used to investigate the effect of F0 on the auditory representation of vowels. In particular, the predictions of Traunmüller's (1981) hypothesis concerning vowel normalization are tested. Section 3.3 reports the results of four simulations of the mixed-F0 condition of Experiment 1. The simulations implement different approaches to perceptual contrast.

An Auditory Model

The auditory model described here incorporates some of the frequency and amplitude nonlinearities found in psychophysical studies of auditory processing, and a spectral integration stage which simulates Chistovich et al.'s (1979) 'centre of gravity' hypothesis. The work of Schroeder, Atal and Hall (1979) and Bladon and Lindblom (1981) formed the foundation of the approach I adopted here. In particular, the use of spectral integration is an important hypothesis concerning the way in which hearers estimate the broad features of the spectral envelopes of speech sounds as those broad features relate to the vocal tract transfer function. It is important to keep in mind, however, that this model is only a *rough* implementation of some hypotheses concerning the human auditory treatment of speech sounds.

The first stage involves the calculation of the magnitude spectrum of a Hamming window of speech samples. In the second stage, the magnitude spectrum is conditioned by a bank of filters. Following Patterson (1976), the equivalent rectangular bandwidth (BW_{ER}) of the filters is given by (1), and the auditory filter shape is given by (2). In these equations f_0 refers to the center frequency of the filter in Hz.

$$10 \log_{10} BW_{ER} = 8.3 \log f_0 - 2.3 \tag{1}$$

$$|H(\Delta f / f_0)^2| = \exp[-\pi(\Delta f / f_0 BW_{ER})^2] \tag{2}$$

These filter shapes are Gaussian approximations to the filter shapes determined by psychoacoustic masking studies (see Moore and Glasberg, 1983). Filter functions were calculated at intervals of 0.2 Bark for the range 0.2 to 19 Bark (18 to 4884 Hz). The output of each filter (A_j) is determined by (3), where n is the number of terms in the filter, W_{ij} is the i^{th} term of the j^{th} filter, and S_i is the spectral magnitude at the frequency corresponding to the i^{th} term of the j^{th} filter. The sum of the products of the magnitude spectrum and the filter weights is normalized by the total of the filter weights because the number of terms in each filter is a function of center frequency.[3]

$$A_j = \sum_{i=0}^{n} S_i W_{ij} / \sum_{i=0}^{n} W_{ij} \tag{3}$$

Next, a stored equal-loudness contour (Figure 3) is applied to the spectrum. Each frequency location in the filtered spectrum is attenuated by the amount indicated in the equal loudness contour, with low frequency components being attenuated more than others.

Insert Figure 3 about here

Figure 4a shows the Fourier transform of a synthetic [ɑ] with an F0 of 120 Hz. Figure 4b shows the same spectrum after passing it through the filter bank and applying the equal-loudness contour. Note that in the filtered spectrum, all but the two lowest harmonics

---

[3] This is necessary because the bandwidths of the filters increases as a function of the center frequency. The samples in the Fourier transform are linearly spaced in frequency, therefore as the bandwidths of the filters increases, the number of samples under the filter window increases.

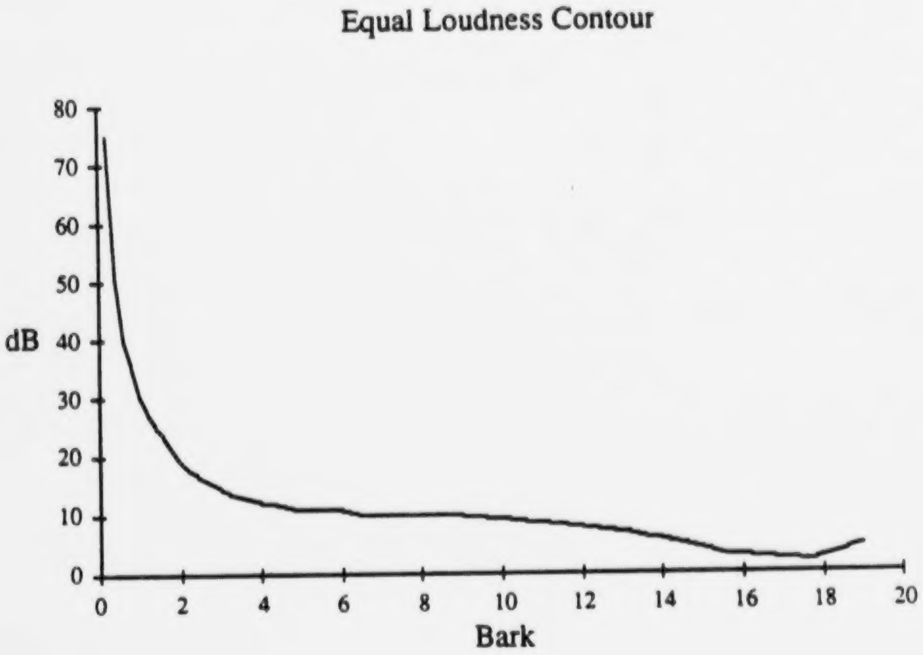


Figure 3. Equal loudness contour, derived from Fletcher and Munson (1933). Ordinate indicates degree of attenuation at each frequency (abscissa).

are smeared together. This is a result of the increasing bandwidths of the auditory filters; frequency resolution decreases as frequency increases. Note also that the regions for F1 and F2 are expanded (as compared with the Fourier transform) and that the higher frequency regions are compressed. Both of these observations are also true of displays of auditory nerve responses to vowel sounds (Sachs and Young, 1979).[4] The model, at this stage of processing, captures some basic properties of peripheral auditory processing.

Insert Figure 4 about here

The third stage of the model involves sliding a window of integration across the spectrum. This stage is an implementation of Chistovich et al.'s (1979) hypothesis that, in vowel perception, spectral components over a fairly large spectral range are integrated into a single 'centre of gravity'. There is no evidence of spectral integration of this sort in neural responses in the auditory pathways of animals (even in the tonotopically organized regions of the cortex, see Pickles, 1988); therefore, if there is a stage of spectral integration in human auditory processing, it is most likely a speech specific, central auditory process (see Chistovich, 1985 and Traunmüller, 1982).

In this model, the Riemann sum (4) over a portion of the spectrum served as an approximation to the definite integral for that spectral region. In (4), Δx_k was 0.2 Bark for all k (the interval between samples in the filtered spectrum) and $f(t)$ was the filtered, loudness-equalized spectrum (Figure 4b). The program calculated Riemann sums over successive windows in the filtered spectrum to produce a power density spectrum. The y dimension for each frequency bin of the power density spectrum was the Riemann sum over a 2.2 Bark window centered on that frequency, and separate sums were calculated at intervals of 0.2 Bark. The sums are expressed in dB normalized to the RMS amplitude of the original waveform in order to preserve relative amplitude differences across speech samples. Experimentation with previous versions of the model indicated that integration over a window of 2.2 Bark resulted in a single spectral peak between F1 and F2 when they were within three Bark of each other. Wider integration windows resulted in the merger of formants which were separated by more than 3 Bark.

$$\sum_{k=1}^{n} f(t_k)\Delta x_k \qquad (4)$$

---

[4] Young and Sachs, 1980 emphasized, also, the fact that firing rate saturates, and consequently that spectral specificity is lost at moderate amplitudes. They suggest that temporal measures such as localized phase locking must also be involved in auditory frequency resolution. Thus, the similarities between the filtered, loudness-equalized spectrum (based on psychophysical studies of hearing) and published displays of mean firing rate or Average Localized Synchronized Rate (ALSR), for that matter, are at best merely suggestive.

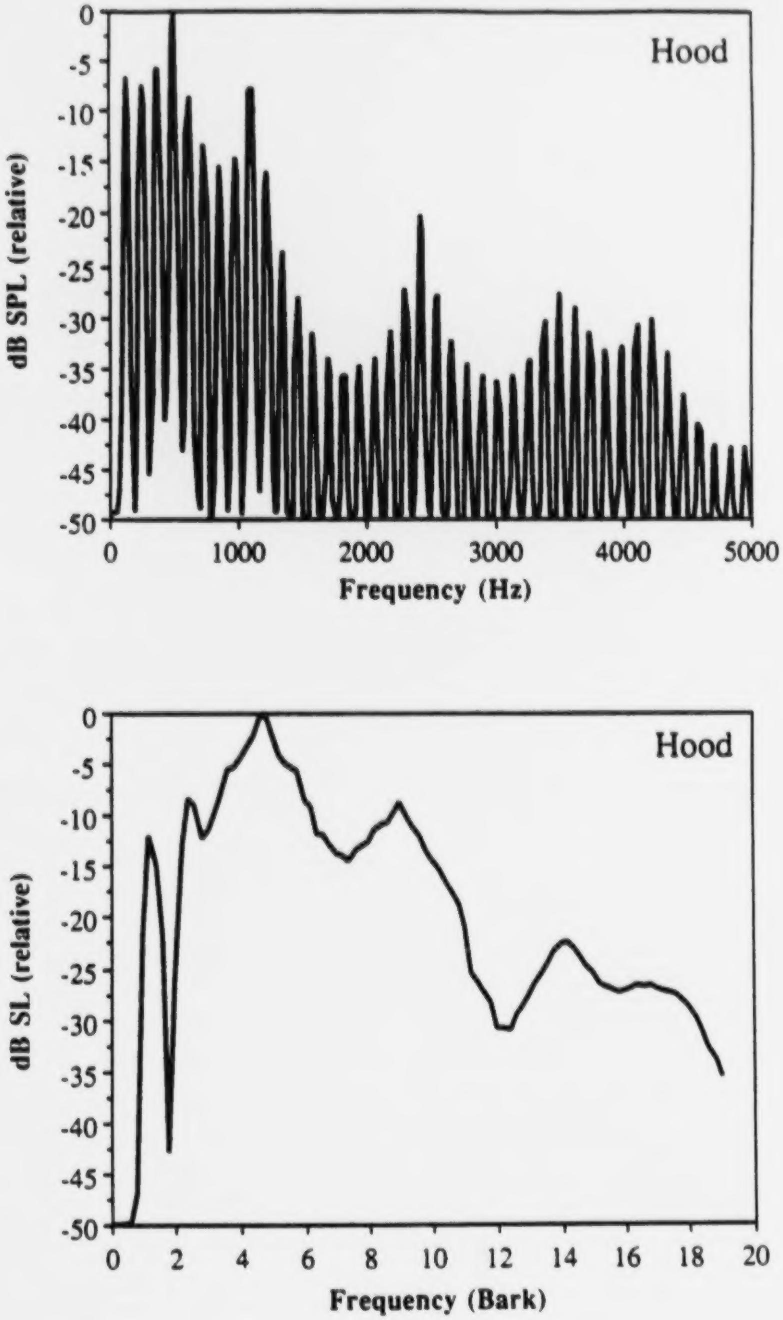


Figure 4. Spectra of the vowel [ʊ] of "hood" (a) The Fourier transform. (b) Filtered, loudness-equalized spectrum.

Because the spectra generated by this model are based on a *rough* attempt to implement some hypotheses concerning the auditory processing of speech, I will refer to spectra produced by the model as "auditorily-based" (AB) spectra. The following section examines the effect of F0 in AB spectra, and section 3.3 reports the results of some simulations of the mixed condition of Experiment 1.

The Effect of F0 in Simulated Auditory Spectra

If F0 normalization can be attributed to the 'centre of gravity' effect, as suggested by Traunmüller (1981), we expect that the model described above will produce spectra in which, as F0 increases, the spectra of "hud" tokens become more like "hood", and as F0 decreases, the spectra of "hood" tokens become more like "hud".

Consider first the AB spectra of the "hood-hud" continuum with low F0 (Figure 5a). There appear to be two primary differences between "hood" and "hud". First, the frequency locations of spectral peaks are lower for "hood" and, second, the amplitude of the spectrum above about 6 Bark is higher for "hud". Figure 5b shows the AB spectra of the continuum with high F0. In these spectra, the first peak occurs at the same frequency throughout the continuum. The difference between "hood" and "hud" for these tokens is mainly in the amplitude of the components around 8 to 10 Bark and the amplitude of the first peak itself ("hood" has the higher amplitude first peak).

Insert Figures 5 & 6 about here

Figure 6 illustrates the effect of F0 on the AB spectra of these vowels. In this figure, we compare the AB spectra of vowels which have identical formant values and different F0 values. The top panel shows the effect of F0 on the spectra of the "hud" endpoint of the continuum. As F0 increases, the frequency of the first spectral peak decreases. This is exactly what an auditory model of F0 normalization would predict. The bottom panel of the figure shows the AB spectra of the "hood" endpoint of the continuum. Here there is no correlation between F0 level and the location of the first spectral peak. This finding is especially troublesome for the auditory approach because with its lower F1 we would expect "hood" to be more sensitive to F0 than "hud".

These data do not lead to the conclusion that as F0 increases AB spectra of "hud" become more "hood"-like, or that as F0 decreases AB spectra "hood" become more "hud"-like (Traunmüller, 1981), but rather that, as F0 increases, the AB spectrum is more and more determined by the harmonics of the fundamental. This is true for AB spectra as well

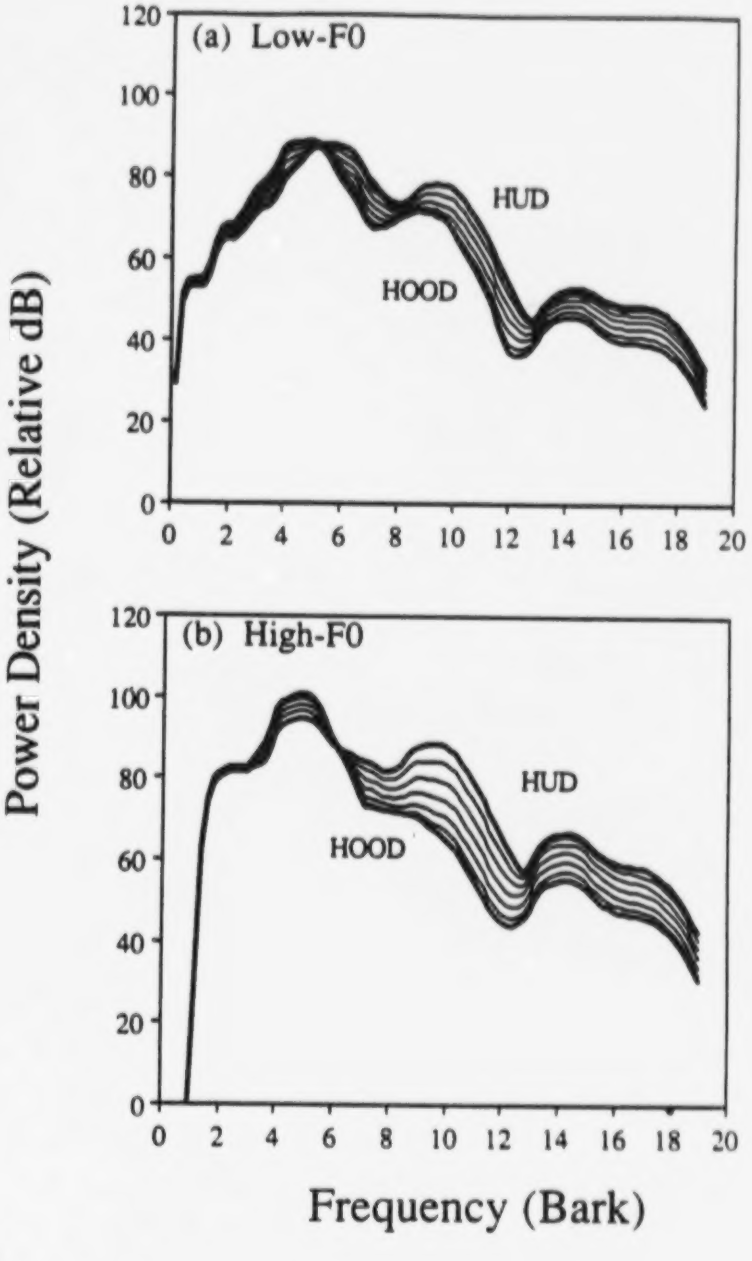


Figure 5. AB spectra of the vowels in the low and high F0 "hood-hud" continua used in the Experiment 1. (a) Low F0 continuum. (b) High F0 continuum.

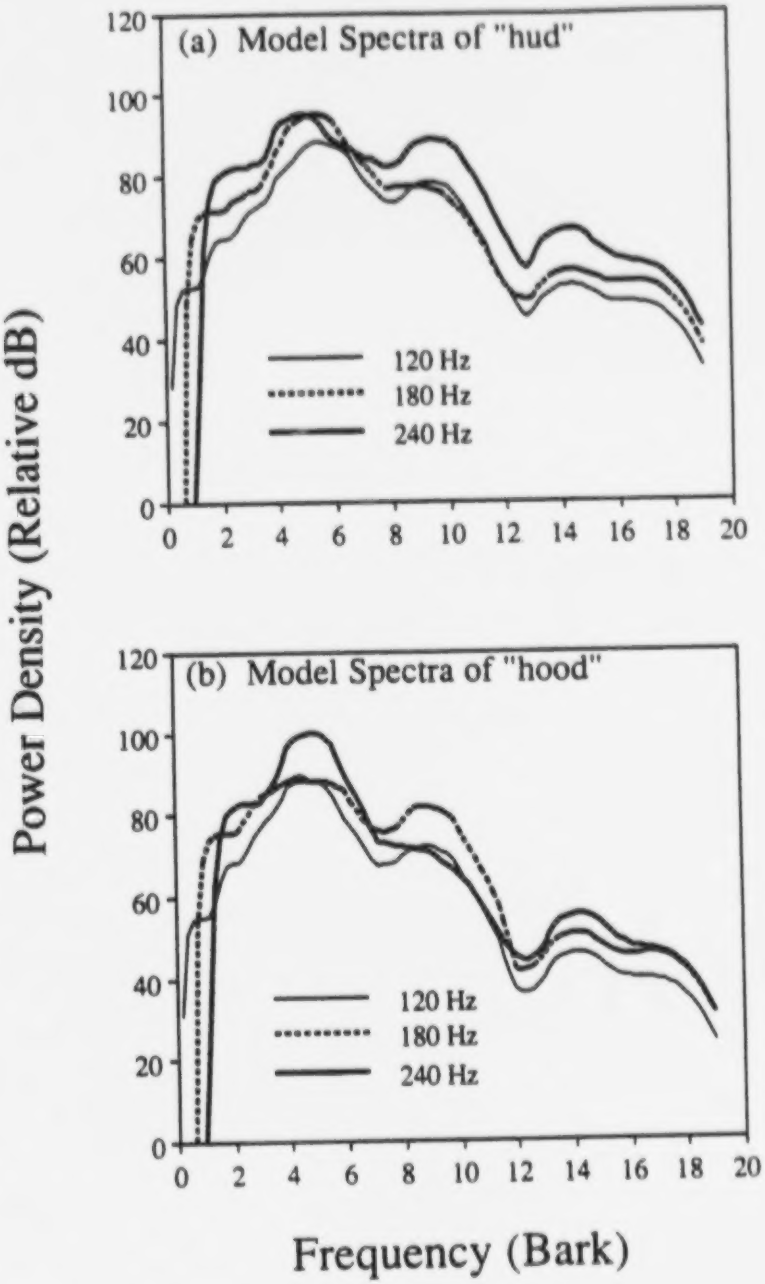


Figure 6. (a) AB spectra of "hud" with F0 of 120, 180, and 240 Hz. (b) AB spectra of "hood" with F0 of 120, 180, and 240 Hz.

as for simple Fourier transforms. Consider the frequency locations of the harmonics relative to the formant peaks. A vowel with a fundamental frequency of 120 Hz will have harmonics at 1.29, 2.53, 3.7, 4.79, 5.78 and 6.83 Bark. The F1 of the synthetic "hood" was 4.74 Bark and of "hud" 5.6 Bark. When F0 is low, the harmonics are closely spaced and, thus, there are harmonics near the F1 of both "hood" and "hud". When the fundamental frequency is 180 Hz, there are harmonics at 1.9, 3.7, 5.3 and 6.69 Bark. So, the third harmonic is close to the F1 of "hud", but F1 falls between harmonics in "hood". Note the broad, flat peak in the AB spectrum of "hood" synthesized at 180 Hz (Figure 6b). When the fundamental frequency is 240 Hz, there are harmonics at 2.53, 4.79 and 6.83 Bark. In this case, the only harmonic in the region of F1 (for these two vowels) is the second harmonic. In the AB spectrum of "hood" the second harmonic and F1 are almost identical (note the amplitude of the peak in Figure 6b). While in the AB spectrum of "hud", the second harmonic is still closer to F1 than are the other harmonics, but the two are not aligned (again the amplitude of the peak seems to reflect this (Figure 6a and Figure 5b). Thus, in the case of vowels with high F0, the shape of the AB spectrum (like other spectral representations) is determined by both the harmonics of the fundamental and by the vowel formants. This investigation of hypothetical perceptual representations of vowel spectra offers no support for Traunmüller's (1981) hypothesis that F0 normalization is the result of the auditory integration of F1 and F0 in a single 'centre of gravity'. Of course, the validity of this conclusion is dependent upon the validity of the model. The fact that the model produces spectra with a single spectral peak between F1 and F2 when they are within 3 Bark of each other suggests that it does capture Chistovich et al.'s (1979) proposal, and, thus, is appropriate for testing Traunmüller's hypothesis.

Model Studies of Contrast in Vowel Perception

The studies reported in this section evaluate the success of four different models of perceptual contrast in accounting for the data from the mixed condition in Experiment 1 (Figure 2b). AB spectra of the vowel tokens used in Experiment 1 above (see Figure 5) served as the input representations in the models of contrast. The first is an implementation of Crowder's (1981) model. The second and third models implement an auditory figure/ground model of vowel contrast. AB spectra in these models were normalized before being subjected to contextual influence (following a suggestion by Fox, 1985, p. 1557). The second model includes an implementation of Bladon et al.'s (1984) spectral shifting approach to vowel normalization and the third model includes an implementation of Gerstman's (1968) range normalization approach. The fourth model incorporates a talker contrast model implied by Mullennix et al.'s (1989) suggestion that hearers adjust to different talkers.

Luce's (1959) choice rule forms the basis of the decision component in each of these models. This rule is defined by (5). In this formula, S_{ij} refers to the similarity of token i to category j and b_j refers to response bias for category j.[5] S_{ij} was calculated by taking the inverse of the spectral distance between the test spectrum and a spectral template. Distance was calculated using (6) after removing the DC offset for differences in overall amplitude. In this formula, S_i is the AB spectrum of token i and S_j is the AB spectrum of template j. The interval *ab* excludes the lowest two Bark and the highest two Bark because incomplete integration windows spanned these edge samples. (6) is similar to the spectral distance metrics used by Plomp (1976, p.95) and Bladon and Lindblom (1981). However, where they used the Euclidian or the city block measures of spectral distance, the mean squared distance was used here. When the inverse of this measure of distance was used as S_{ij} in (5), the resulting response functions were very similar to those found in the blocked condition of Experiment 1 (Figure 2a). This similarity is an important starting point for the simulations of the mixed condition (Figure 2b).

---

[5] b_j was held fixed at 0.5 in the first models and was allowed to vary in the last model.

$$P(R_{ji}|T_i) = b_j S_{ij} / \sum_{j=1}^{2} b_j S_{ij} \tag{5}$$

$$D_{ij} = \left(\sum_{x=a}^{b} (S_i(x) - S_j(x))^2 \right) / (b - a) \tag{6}$$

Crowder's Model of Contrast. In the implementation of Crowder's model of contrast, AB spectra from the high and low "hood"-"hud" continua were classified based on a comparison with stored templates for "hood" and "hud". The average of the AB spectra of the high- and low F0 endpoints of the continua served as templates in this model. The model classified each vowel in the context of every other vowel (both within and across continua), after the context spectrum had been attenuated by a certain proportion (the **decay** parameter), and then subtracted from the test spectrum. If the context spectrum has very little energy at a particular frequency, then the test spectrum will remain relatively unchanged at that frequency. However, if the two spectra have peaks in about the same location in frequency, then the peak of the test spectrum will be reduced (to an extent determined by the decay parameter) as a result of context. The value of the decay parameter which provided the best fit to the data Experiment 1 (Figure 2b) was estimated by the method of least squared error. No value of the decay parameter provided a very close fit to the data. The RMS error (which is in the same units as the data, in this case, percent "hood" responses) of the best fit obtained was 35.4 (Table 3).

This simulation indicates that a model of vowel contrast along the lines of that proposed by Crowder (1981) does not account for the contrast effect found in the Experiment 1. This may be an indication that different talkers should be viewed as occupying different "channels" in auditory memory and that we should not expect recurrent lateral inhibition to play a role in contrast when two different voices are involved.

Insert Table 3 about here

Auditory Figure/Ground Contrast. In the auditory figure/ ground models of vowel contrast, the probability of a "hood" response for context items influenced the probability of a "hood" response on the current item. If the immediately preceeding item was very much like "hood", then an ambiguous item will be more likely to be identified as "hud" than if the context item was a good example of "hud". In this model of vowel contrast (7), the adjusted probability of a "hood" response ($\Pi_n(hood)$) was equal the base probability of the current item (defined by (5)) mulitplied by the ratio of the base probability of the current item and the base probability of the immediately preceding context item, with resulting probability values truncated to the range 0 to 1. Each token served as a context for every other token in computing the average probability of a "hood" response.

$$\Pi_n(hood) = P_n(hood) * (P_n(hood)/P_{n-1}(hood)) \qquad (7)$$

Prior to the calculation of context effects, the AB spectra were normalized using an implementation of one of two different approaches to vowel normalization. One model used an implementation of Bladon et al.'s spectral shifting model of normalization. In this approach to normalization, AB spectra of vowels with high F0 were shifted down on the Bark scale and then compared with AB spectral templates appropriate for a male speaker. AB spectra of steady-state tokens synthesized using the Peterson and Barney (1952) average formant and F0 values for male "hood" and "hud" served as vowel templates in this model. The dialect of the speakers in Peterson and Barney's study was similar to that of the subjects in Experiment 1 so the use of these values is appropriate. The second implementation of the auditory figure/ground model used a form of range normalization (Gerstman, 1968). In this implementation of vowel normalization, the choice of templates depended on F0. If F0 was low, the Peterson and Barney average male templates were used. If F0 was high, the templates were derived from the Peterson and Barney average vowel formants and F0 for females.

In the spectral shifting model, the degree of shift was a free paramcter. The degree of spectral shift, estimated by the least squared error method, was 0.8 Bark. This corresponds quite closely to the value used by Bladon et al. (1984). There were no free parameters in the range normalization model. Although these two approaches to vowel normalization classify the continua in quite different ways in the absence of any contrast effect (the spectral shifting model classifies both continua more consistently), they provide virtually identical results when used in the vowel contrast model. RMS error of both the spectral shifting model and the range normalization model was 19.8. Both of these models provide a better fit to the data than does the lateral inhibition model, but the predictions are still pretty

Table 3

RMS data and parameter estimates for the model studies.

Model	RMS Error	Parameter Values
Lateral Inhibition	35.4	decay = 0.998
Figure/ground Contrast		
Spectral Shifting	19.8	shift = 0.8
Range Normalization	19.8	
Figure/ground with Lateral Inhibition		
Spectral Shifting	10.3	shift = 0.8 decay = 0.3
Range Normalization	10.3	decay = 0.3
Talker Contrast	8.9	bias = 0.75

rough. However, when the lateral inhibition context effect is included in these models, the degree of fit improves considerably. RMS error was 10.3 for both the spectrum shifting and range normalization models. Predicted identification functions and the data obtained in the mixed condition of Experiment 1 are shown in Figure 7a.

Lateral inhibition affects the spectral representation of the stimulus before it is compared with a spectral template, and figure/ground contrast affects the decision rule used to classify the stimulus after spectral similarity has been calculated. It is, therefore, reasonable to expect that both lateral inhibition in auditory memory and figure/ground contrast could be involved in vowel contrast effects since they occur at different stages of processing. Note, however, that this conclusion, unlike the one reached in the previous section, suggests that different voices are interacting in auditory memory (perhaps across "channels").

Insert Figure 7 about here

Talker Contrast. The final model implements a model of talker contrast. If hearers adjust their expectations for vowel quality relative to the perceived identity of the talker, as suggested indirectly by Mullennix et al. (1989) and more directly by Johnson (in press), we could suppose that these expectations would result in a contrast effect which depends on perceived speaker characteristics rather than vowel quality. In other words, the contrast effect observed in Experiment 1 can be considered a talker contrast effect rather than a vowel contrast effect.

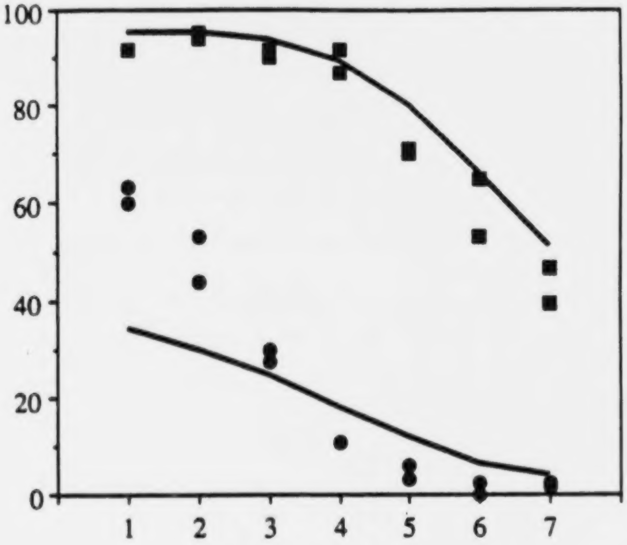
In the implementation of this view, the response bias factor (b_j) in the decision rule (5) varied as a function of the F0 of the item. Low F0 items had to be quite similar to "hood" in order to be classified as "hood", while high F0 items had to be quite similar to "hud" in order to receive that label. This reflects the hypothesis of an adjustment-to-talker model of vowel normalization that the criteria for vowel classification are a function of the perceived identity of the talker. There was a single free parameter in the model. This parameter (the **bias** parameter) functioned as b_{hood} when F0 was high and b_{hud} when F0 was low. Bias toward the other category in each case was equal to $1 - bias$.[6]

Response probabilities predicted by the talker contrast model are shown in Figure 7b. As the figure shows, this model of contrast is also quite accurate in predicting the data of Experiment 1 (the RMS error was 8.9).

---

[6] The spectra derived from the Peterson and Barney (1952) average formant values for male speakers were used as templates in this model.

(a) Vowel Contrast Predictions



(b) Talker Contrast Predictions

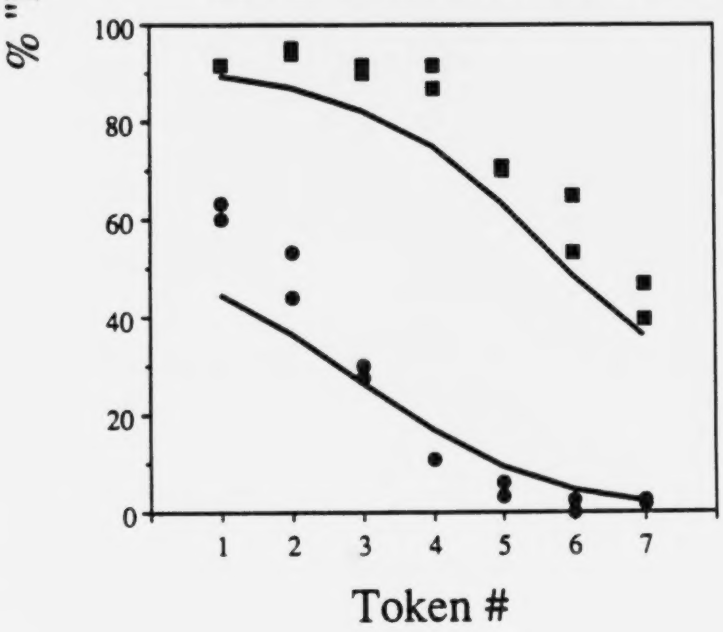


Figure 7. Results of simulations of the mixed condition of Experiment 1 using (a) a post-normalization vowel quality contrast model, and (b) a talker contrast model. Predictions are the solid lines and the data from Experiment 1 (see Figure 2b) are plotted with filled squares (high F0) and circles (low F0).

Summary

The findings of the model studies are summarized below:

(1) The hypothesis that vowel normalization is a consequence of auditory processing was not supported. In spectra generated by a model which incorporates two stages of filtering, as in other forms of frequency analysis, when F0 increases, the shape of the spectrum becomes more and more dependent upon the harmonics of the fundamental frequency.

(2) Crowder's (1981) model of contrast cannot account for the results of Experiment 1. However, in combination with an auditory figure/ground contrast model it does help provide a close fit to the data. Unresolved is the question of whether vowels produced by different talkers should be seen as interacting in auditory memory.

(3) Range normalization (as implemented here) and normalization by means of spectral shifting (Bladon et al., 1984) give identical fits to the data of Experiment 1 when they are used to provide the input to a figure/ground contrast mechanism.

(4) Two types of contrast provide good fits to the data - (1) auditory figure/ground contrast coupled with recurrent lateral inhibition and (2) talker contrast. The first involves contrast in vowel quality, the second in perceived talker identity.

Experiment 2

The two types of contrast which best account for the mixed-F0 data of Experiment 1 make very different predictions for a cross-series anchoring experiment. If the contrast effect observed in Experiment 1 occurred primarily at the level of vowel quality, we predict that the perceived vowel quality of the anchor token will be the dominant factor in cross-series anchoring. Conversely, if the contrast effect found in Experiment 1 was the result of a talker contrast process, we predict that the vowel quality of the anchor will be of less importance than the perceived identity of the talker, and thus, that the same direction of boundary shift will be produced by anchors of different vowel quality.

Experiment 2 is a test of these predictions. The stimuli which were used in Experiment 1 were presented in a cross-series anchoring experiment. Subjects heard the items of one of the two continua (low or high F0) and then heard those same stimuli randomly intermixed with multiple occurrences of an anchor stimulus drawn from the other continuum.

Method

Subjects. Thirty-eight undergraduate students (10 male, 28 female) at Indiana University participated in the experiment for partial course credit in an introductory psychology course. All were native speakers of American English who had never experienced any speech or hearing disorders.

Materials. This experiment employed the same stimuli which had been used in Experiment 1.

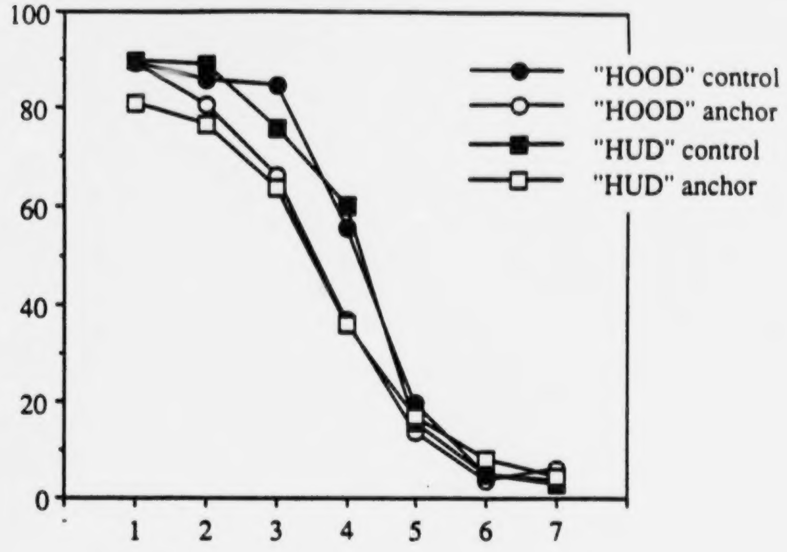
Procedure. Subjects were randomly divided into four groups (two groups of 11 and two groups of 8). Each group was presented with a randomized list containing 20 repetitions of each of the tokens from either the high F0 or low F0 continuum in a control condition and then with those same tokens randomized with 60 occurrences of an anchor token from the other continuum. Thus, there were 140 trials in the control condition and 200 trials in the anchor condition. The first group of 11 subjects heard the low F0 continuum in the control condition and the low F0 continuum with 60 occurrences of token 1 from the high F0 continuum in the anchor condition. Group two (11 subjects) also heard the low F0 continuum, but with token 7 from the high F0 continuum as an anchor. Groups three and four (8 subjects in each group) responded to the tokens of the high F0 continuum with tokens 1 and 7 (respectively) of the low F0 continuum as anchors. The equipment used to run the experiment was the same as that used in Experiment 1.

Results

The results of Experiment 2 are shown in Figures 8 and 9. Figure 8 shows the identification responses plotted by token number. Panel (a) shows the data for the low F0 continuum for both the "hood" and "hud" anchor groups. The data presented in this panel were analyzed in a three-factor, repeated-measures analysis of variance. Factors were: CONDITION (control versus anchor), ANCHOR ("hood" versus "hud"), and TOKEN. There was (predictably) a main effect for TOKEN [$F(6,60)=199.27$, $p<0.001$]. More to the point, there was also a main effect for CONDITION [$F(1,10)=14.4$, $p<0.01$]. In the control condition, the average percent "hood" response was 48.5%, while in the anchor condition this was reduced to 41.6%. The only other effect which reached significance was the CONDITION by TOKEN interaction [$F(6,60)=8.52$, $p<0.001$]. The effect of anchoring was to shift the phoneme boundary rather than producing a global change in probability of a "hood" response.

Insert Figures 8 & 9 about here

(a) Low-F0 Continuum: High-F0 Anchors



(b) High-F0 Continuum: Low-F0 Anchors

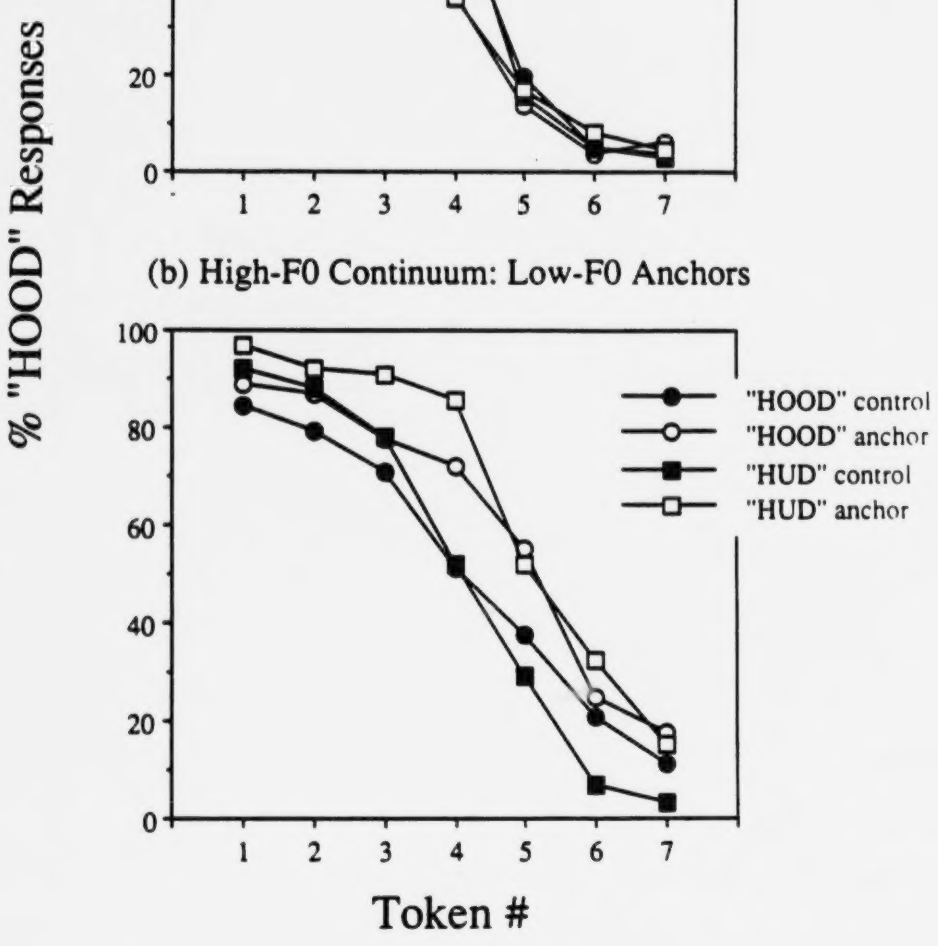
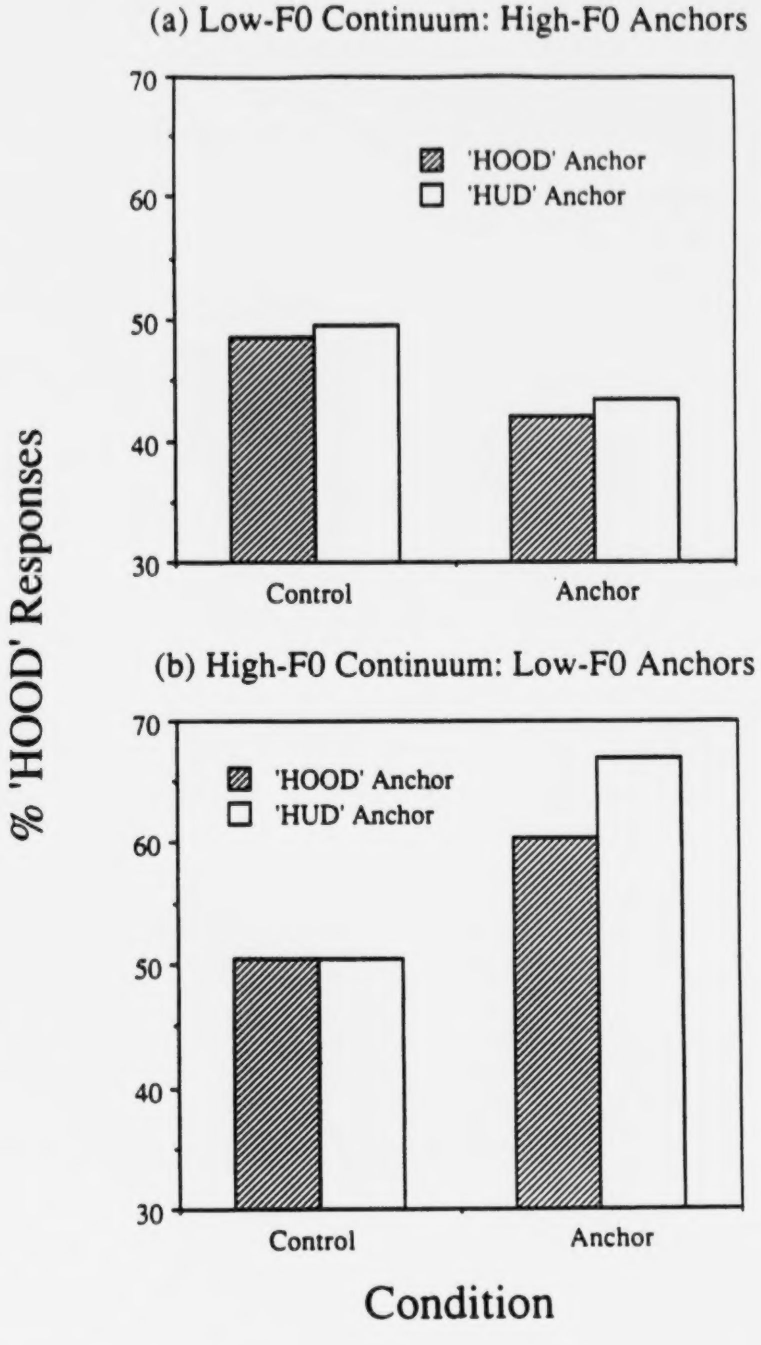


Figure 8. Results of Experiment 2. Identification functions for (a) the low F0 "hood"-"hud" continuum, control conditions (solid points) and the anchor conditions (open points), and (b) the high F0 "hood"-"hud" continuum.


(a) Low-F0 Continuum: High-F0 Anchors

(b) High-F0 Continuum: Low-F0 Anchors

Figure 9. Results of Experiment 2 averaged over tokens. Percent "hood" responses to (a) the low F0 "hood"-"hud" continuum, and (b) the high F0 "hood"-"hud" continuum.

Figure 8b shows identification functions for the high F0 continuum. These data were also analyzed in an ANOVA with factors: CONDITION, ANCHOR, and TOKEN. The same three statistical effects were significant in this analysis. There were main effects for TOKEN [$F(6,42)=116.2, p<0.001$] and CONDITION [$F(1,7)=22.42, p<0.01$], and the CONDITION by TOKEN interaction was significant [$F(6,42)=3.91, p<0.01$]. The CONDITION by ANCHOR interaction approached significance, but was not reliable [$F(1,7)=2.08$, $p=0.19$]. This interaction is most visible in Figure 8b as the difference between the "hood" and "hud" anchor conditions for tokens 3 and 4.

Discussion

These results quite clearly conform to the predictions of the talker contrast model. One concern should be addressed, though. The "anchor" tokens in this experiment were not uniformly identified as "hood" or "hud" in Experiment 1. In particular, the "hud" endpoint token was identified (in the mixed condition) as "hud" only slightly more than 50% of the time when it had high F0. Also, the "hood" endpoint token of the low F0 continuum (again in the mixed condition) was identified as "hood" only 60% of the time. Similarly, in the present experiment, when the anchor stimulus had high F0, the "hood" anchor was identified as "hood" 88.7% of the time, while the "hud" endpoint anchor was identified as "hud" only 46.3% of the time. When the anchors had low F0, the "hood" endpoint anchor was identified as "hood" 49.8% of the time and the "hud" endpoint anchor was identified as "hud" 97.4% of the time. Thus, these tokens do not satisfy the requirements of a model which relies on contrasting vowel quality.

However, not all of the subjects identified the anchor tokens in this manner. There were three subjects in the high F0 "hud" anchor group who reliably (i.e. more than 75% of the time) identified the anchor as "hud". An analysis of variance of the data from these subjects also supported the talker contrast hypothesis. Factors were TOKEN and CONDITION. There was a reliable interaction between these factors which indicated a boundary shift [$F(6,12)=3.27$, $p<0.05$]. The direction of this boundary shift was the same as the shift found in the overall analysis of variance. When the anchor had high F0 and the continuum low F0 the subjects tended to label ambiguous stimuli more as "hud" than they did in the control condition.

Of the subjects in the low F0 "hood" anchor group, there were three who reliably labelled the anchor as "hood". An analysis of the data from these three subjects revealed a similar trend. In this analysis the CONDITION main effect approached significance [$F(1,2)=4.38$, $p=0.17$]. The direction of this trend conformed with the overall analysis. When high F0 items were identified in the context of low F0 anchors (even anchors which were identified as "hood"), subjects tended to respond "hood" more frequently than they did in the control condition.

Note also that if the overall data were determined by vowel quality contrast, we would predict that when ambigous anchors were used there would have been no boundary shift. The fact that two stimuli which had quite ambiguous vowel qualities produced boundary shifts just as large (and in the same direction) as unambiguous vowels is further evidence that the boundary shift observed here is not the result of a contrast in vowel quality. If anchoring occurred at the level of perceived vowel quality, we would predict that when the anchor is "hood" subjects will use the "hud" label more often, or that when the anchor is "hud" subjects will use the "hood" label more often. This prediction is not borne out in these separate analyses nor in the overall analysis. Rather, the data conform to the predictions of the talker contrast model. When F0 of the anchor is high, perception of low F0 tokens is shifted toward "hud", and when the F0 of the anchor is low, perception of high F0 tokens is shifted toward "hood".

Fox (1985) conducted a very similar cross-series anchoring experiment and got very different results. The main difference between his experiment and the present one concerned the stimuli. Fox used two continua from "hid" to "head". In one case, the formant range of the continuum was appropriate for a male talker, in the other for a female talker. The tokens with a relatively high formant range were synthesized with high F0 and the tokens with a relatively low formant range were synthesized with both high and low F0. On the other hand, the stimuli used in this experiment formed a continuum from "hood" to "hud" and occupied a formant range which was ambiguous between male and female average values. Both the difference between front and back vowels and overall formant ranges contribute to the discrepancy of results. First, in the back vowel continuum used here F1 and F2 are positively correlated across the continuum. F1 and F2 both increase from the "hood" endpoint to the "hud" endpoint. In Fox's front vowel continua, F1 and F2 were negatively correlated. F1 increased from "hid" to "head" while F2 decreased from "hid" to "head".[7] If hearers expect generally higher formants when F0 increases, it is not clear how a continuum in which F1 and F2 are negatively correlated would be handled perceptually. There is some evidence that F1 is more affected by normalization than is F2 (Ainsworth, 1975), but it is also likely that when information from F2 contradicts information from F1, the F1 information will be less useful than when F1 and F2 are correlated. Second, the continuum used here spanned formant ranges which were ambiguous between male and female values. This, coupled with the correlation of F1 and F2, meant that this continuum was very sensitive to a normalization effect; when experimentally manipulated factors influenced perceptual normalization these manipulations were easily observable in subjects' responses to the vowels from the continuum. It is not clear how a talker contrast effect could have produced a shift of identification in Fox's (1985) "hid"-"head" continua.

---

[7] It should also be noted that F1 and F2 were also positively correlated in the stimuli used by Fujisaki and Kawashima (1968).

Conclusion

The results of this investigation provide indirect evidence for a talker contrast effect. Research is currently under way to test for the existence of such an effect more directly. If the interpretation given above is correct, and a contrast at the level of perceived talker identity is actually taking place, then only one view of vowel normalization remains tenable. In both of the experiments reported here, vowel identification functions were influenced by context. The effect of context was to increase (or cause?) the vowel normalization effect (i.e. tokens with high F0 had to have higher formant values to be identified as "hud" and tokens with low F0 had to have lower formant values to be identified as "hood"). The data of Experiment 2 suggest that the influence of context is at the level of talker quality and not vowel quality. Therefore, we conclude that the vowel normalization effect is influenced by talker quality, or, more generally, that perceptual vowel normalization makes reference to percieved talker identity. Of course, it is necessary to point out that this conclusion is based on vowel identification performance in response to only one vowel continuum ("hood" - "hud"). Therefore, the general validity of these results for other vowel contrasts remains to be shown. Assuming that these results are generally valid for other vowels and other languages, they suggest that the algorithmic approach to vowel normalization which is exemplified by Gerstman (1968), Labonov (1971), Nearey (1978), Disner (1980), Sussman (1986), Syrdal and Gopal (1986), Miller (1989) and others, has left out one crucial variable. The information that hearers use to evaluate vowel quality includes not only acoustically available information (such as vowel spectrum and F0), but also computed information about the person doing the talking.

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgements. In Fant, G. & Tatham, M., editors, *Auditory analysis and perception of speech*. Academic Press, London.

Bladon, R., Henton, C., & Pickering, J. (1984). Towards an auditory theory of speaker normalization. *Language and Communication*, **4**, 59–69.

Bladon, R. & Lindblom, B. (1981). Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America*, **69**, 1414–1422.

Chistovich, L., Sheikin, R., & Lublinskaja, V. (1979). 'Centres of Gravity' and spectral peaks as the determinants of vowel quality. In Lindblom, B. & Öhman, S., editors, *Frontiers of speech communication research*. Academic Press, London.

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America*, **77**, 789–805.

Cooper, W. E. (1974). Adaptation of phonetic feature analyzers for place of articulation. *Journal of the Acoustical Society of America*, **56**, 617–627.

Crowder, R. (1981). The role of auditory memory in speech perception and discrimination. In Myers, T., Laver, J., & Anderson, J., editors, *The cognitive representation of speech*. North-Holland, New York.

Delattre, P., Liberman, A., Cooper, F., & Gerstman, L. (1952). An experimental study of the acoustic determinants of vowel colour. *Word*, **8**, 195–210.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253–261.

Eimas, P. D. (1963). The relationship between identification and discrimination along speech and nonspeech continua. *Language and Speech*, **6**, 206–217.

Eimas, P. D., Cooper, W. E., & Corbit, J. D. (1973). Some properties of linguistic feature detectors. *Cognitive Psychology*, **4**, 99–109.

Eimas, P. D. & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Perception and Psychophysics*, **13**, 247–252.

Fletcher, H. & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, **5**, 82–108.

Fox, R. (1985). Auditory contrast and speaker quality variation in vowel perception. *Journal of the Acoustical Society of America*, **77**, 1552–1559.

Fry, D., Abramson, A., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, **5**, 171–189.

Fujisaki, H. & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE-AU*, **16**, 73–77.

Fujisaki, H. & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, pages 67–73.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE-AU*, **16**, 78–80.

Johnson, K. (1989). Higher formant normalization results from auditory integration of F2 and F3. *Perception and Psychophysics*, **46**, 174–180.

Johnson, K. (in press). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America.*

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971–995.

Labonov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, **49**, 606–608.

Luce, R. (1959). *Individual choice behavior.* Wiley, New York.

Miller, J. (1989).Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114–2134.

Miller, R. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America*, **25**, 114–121.

Moore, B. C. J. & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, **74**, 750–753.

Mullennix, J., Pisoni, D., & Martin, C. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365–378.

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* IU Linguistics Club, Bloomington, Indiana.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, **72**, 407–418.

Parducci, A. (1975). Contextual effects: A range-frequency analysis. In Carterette, E. C. & Friedman, M. P., editors, *Handbook of perception, Vol. II*. Academic Press, New York.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, **59**, 640–654.

Peterson, G. & Barney, H. (1952). Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America*, **24**, 175–184.

Pickles, J. (1988). *An introduction to the physiology of hearing*. Academic Press, London, 2nd edition.

Pisoni, D. (1971). *On the nature of categorical perception of speech sounds*. PhD thesis, University of Michigan.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253–260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory and Cognition*, **3**, 7–18.

Plomp, R. (1976). *Aspects of tone sensation: A psychophysical study*. Academic Press, London.

Sachs, M. & Young, E. (1979). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *Journal of the Acoustical Society of America*, **66**, 470–479.

Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Objective measure of certain speech signal degredations based on masking properties of human auditory perception. In Lindblom, B. & Öhman, S., editors, *Frontiers of speech communication research*. Academic Press, London.

Simon, H. J. & Studdert-Kennedy, M. (1978). Selective anchoring and adaptation of phonetic and nonphonetic continua. *Journal of the Acoustical Society of America*, **64**, 1338–1357.

Slawson, A. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America*, **43**, 87–101.

Summerfield, A. (1971). *Information-processing analyses of perceptual adjustments to source and context variables in speech*. PhD thesis, Queen's University of Belfast.

Summerfield, A. & Haggard, M. (1975). Vocal tract normalization as demonstrated by reaction times. In Fant, G. & Tatham, M., editors, *Auditory analysis and perception of speech*. Academic Press, London.

Sussman, H. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, **28**, 12–23.

Syrdal, A. & Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086–1100.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, **69**, 1465–1475.

Traunmüller, H. (1982). Perception of timbre: evidence for spectral resolution bandwidth different from critical band? In Carlson, R. & Granström, B., editors, *The representation of speech in the peripheral auditory system*. Elsevier Biomedical, Amsterdam.

Young, E. & Sachs, M. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America*, **66**, 1381–1403.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Coronals and the Phonotactics of Nonadjacent Consonants in English[1]

Stuart Davis[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This work was supported, in part, by NIH Training Grant NS-07134-09 to Indiana University. I wish to thank Diana Archangeli, Nick Clements, Mike Hammond, Keith Johnson, Carole Paradis, Jean-Francois Prunet, Doug Pulleyblank, and Moira Yip for their comments on earlier versions of this paper. I am also grateful to Luis Hernandez for his assistance with Lexis on the Symbolics Lisp Machine. A slightly revised version of this paper is to appear in C. Paradis and J.-F. Prunet (eds.), *The Special Status of Coronals*, Foris Publications, Dordrecht.

[2] Now with the Linguistics Department, Indiana University, Bloomington, IN

Abstract

In recent work on phonological features there has been much discussion on coronal consonants. Some researchers have argued based on a variety of phonological and typological considerations that coronal consonants should be represented as lacking place of articulation features at the underlying level of representation. This paper presents English evidence supporting such a view. Specifically, it is shown that English has a Morpheme Structure Condition that disallows sCVC sequences in which the two nonadjacent C's are homorganic. However, this constraint does not apply to coronals since monomorphemes like *state* and *stout* are quite frequent. That coronal consonants are ignored by the constraint can be taken as evidence that English coronal consonants lack place features in underlying representation.

Coronals and the Phonotactics of Nonadjacent Consonants in English

Introduction

Much recent work on underspecification has focussed on the status of coronal consonants. One question this work addresses is whether or not coronal consonants lack the Place Node. (We assume here a theory of Feature Geometry like that proposed in Sagey (1986) and discussed in Davis (1988) in which the Place Node dominates the articulator nodes Labial, Coronal, and Dorsal.) The conclusions reached about the status of coronals is quite varied. Some researchers such as Paradis & Prunet (1989a, 1989b) have concluded that ([+ant]) coronal consonants are different from labials and dorsals in that as a principle of grammar they lack the Place Node. Other researchers, such as Avery & Rice (1989), contend that whether or not coronal consonants in a language lack the Place Node depends on the phonemic inventory of that language, while still other researchers such as Clements (1988) and Mester & Ito (1989) give no special status to coronals with respect to underspecification.

The major reason why these researchers have reached different conclusions is that they have used competing criteria in determining what is underspecified in underlying representation. For example, Avery & Rice (1989) assume that phonemic inventories (and not phonological rules) are relevant for determining what is underspecified. Thus, in their view, the Coronal Node is present in the underlying representation of any two phonemes (in an inventory) that differ only in a feature that is dominated by the Coronal Node (such as [anterior]). Thus in a language like English that has an anteriority contrast between /s/ and /š/, the Coronal Node would be present underlyingly in both sounds, whereas /t/ would not have a Coronal Node since /t/ does not contrast with a corresponding nonanterior coronal stop. In more radical versions of underspecification–e.g. Archangeli (1988), Paradis & Prunet (1989b)–at least some coronal consonants could lack the Coronal Node in underlying representation in spite of a (minimal) contrast with another coronal phoneme. For example, Paradis & Prunet (1989b) contend that Fula coronal consonants that have the feature [+anterior] completely lack the Coronal Node while the corresponding coronal consonants with the feature [-anterior] possess the Coronal Node in underlying representation. Paradis & Prunet base their contention on the fact that vowel spreading and assimilation between nonadjacent vowels occur in the West African language Fula only if the intervening consonant is an anterior coronal. They argue that such consonants must lack the Place Node completely (and consequently the Coronal Node) or else they would not be transparent to vowel spreading and assimilation. Paradis & Prunet show that similar cases of coronal transparency are found in the African languages Guere and Mal. However, in languages where coronals block vowel spreading and assimilation, they propose that coronal consonants are specified for place features early in the derivation. Thus they are able to maintain that ([+anterior]) coronals always lack the Place Node (and thus place of articulation features) in underlying representation.

In more restrictive views of underspecification–Clements (1988) and Mester & Ito (1989)–coronal consonants are not necessarily viewed as having any special status, regardless of the nature of the phonemic inventory. For example, Mester & Ito argue that the Coronal Node must be present underlyingly in all coronal consonants in Japanese (except /r/), even though there are no (phonemic) contrasts between anterior and nonanterior coronals. Their argument for lack of underspecification of coronals is based on palatal prosody in Japanese. This is a process whereby palatalizing certain consonants in the base form of a word adds to the meaning of the word a sense of "uncontrolledness". They contend that which consonants become palatalized is quite predictable: essentially the rightmost coronal consonant of the base (excluding /r/) becomes palatalized, or, in the absence of a coronal consonant, the first consonant of the word becomes palatalized. They argue that this process assumes that coronal consonants cannot be unspecified for place of articulation in underlying representation or else coronal consonants could not be singled out.

Moreover, Yip (1989), Clements (1988), and Mester & Ito (1989) have all employed (to different degrees) an argument against coronal consonants being underspecified based on Morpheme Structure Constraints (MSC's). As Yip (1989) has pointed out, MSC's pertain to underlying representations, and therefore they should have access to underlying specifications only. These authors contend that coronals (especially coronal obstruents) are not treated differently than other consonants (i.e. labial or dorsal) by MSC's. That is, MSC's do not seem to treat coronal consonants as if they were unspecified. This finding has led Mester & Ito to conclude that the coronal place of articulation has no special status, and the finding is also probably a motivating factor behind Clements' proposal that articulator nodes of consonants are always present in underlying representation.

Thus we see that various researchers have reached different conclusions about the underspecification of coronals because they have examined different types of criteria for determining what features (or Nodes) are unspecified. Of these, it is only MSC's that do not seem to treat coronals as special. In this paper I will focus on the MSC argument for (under)specification. I will contend, contrary to Clements (1988) and Mester & Ito (1989), that MSC's can and do treat coronal consonants as special. The evidence to be discussed in this paper comes from English MSC's that pertain to nonadjacent consonants. These MSC's can only be understood if at least some coronals lack the Place Node in underlying representation. This finding, taken together with different MSC's that occur in other languages, provides evidence for a view in which the presence of the Place Node for coronals is a parameterized option.

The organization of this paper is as follows. In Section 2 we show that the need for MSC's cannot be obviated completely as has been argued by Hooper (1975) who contends that MSC's are always reducible to, and thus expressible as, syllable structure constraints (SSC's). It is shown that MSC's are required and that they can be distinguished from SSC's. In Section 3 we show that the English MSC's that hold between nonadjacent consonants do treat coronals differently than noncoronals. It is subsequently argued that this finding provides support for a view of underspecification in which the presence of the Place Node for coronals is a parameterized option.

Distinguishing MSC's from SSC's

Before discussing MSC's that hold between nonadjacent consonants in English, it is important to show that MSC's can be distinguished from SSC's in light of the work of Hooper (1975). Hooper argues against the existence of MSC's altogether. Basically, Hooper contends that all MSC's are expressible as, and so reducible to, Syllable Structure Constraints. Hooper's argument for replacing MSC's with SSC's comes largely from Spanish data. She notes that, in Spanish, there seem to be no constraints on morpheme-final clusters, since final clusters that are impossible in isolated syllables do occur morpheme-finally. Examples of such clusters include *bl* and *pr* which occur in final position in the morphemes *abl* 'speak' and *kompr* 'buy', respectively. In syllable-final position, on the other hand, there are strong constraints on what consonants (and consonant clusters) can occur. Hooper contends that such constraints are missed in an analysis incorporating (only) MSC's. Hooper's example from Spanish, though, does not really argue against MSC's, rather it provides evidence for the necessity of SSC's. That is, there are certain constraints that are best expressed in terms of the syllable.

Hooper does consider the possibility that there are both SSC's and MSC's. She ends up rejecting completely MSC's. One reason that she rejects MSC's is that there would be different MSC's for stems and for suffixes (in Spanish). For example, the sequence *nd* is a possible initial sequence of a suffix (as in the progressive morpheme *ndo*, but it is not a possible stem initial sequence. However, this potential reason for rejecting the existence of MSC's would probably not be relevant if inflectional morphemes (such as the progressive *ndo* in Spanish) are not represented in the Lexicon in the first place. Such a view of inflectional morphemes is argued for by Anderson (1982), Janda (1983), and others who work within the Item-and-Process view of morphology. Anderson, for example, argues that since inflectional morphology is integrated into the syntax, inflectional morphemes could not be listed in the Lexicon. Instead, they are introduced by inflectional rules. Consequently, they may not necessarily display the same sound sequence constraints as the other morphemes of the language.

Hooper does consider one potential MSC for English, but rejects it as being accidental. This is the constraint that rules out morphemes that end in two voiced obstruents. While it is certainly the case that English syllables can end in two voiced obstruents (e.g. *nabbed*, *pigs*), monomorphemes do not–ignoring the uncommon words *adze* and *ides*. In order to defend her position against the necessity of MSC's, Hooper contends that the existence of this constraint is the result of a historical accident and so does not reflect the morpheme structure of English. On the other hand, Kahn (1976:40) has maintained that the constraint really does reflect an MSC. He notes that English nonsense words like [nɛgz] are always interpreted as having two morphemes, and, moreover, English has other possible syllable-final sequences that are not possible morpheme-final sequences (e.g. [ksθs] as in *sixths*). Thus Hooper's contention about the accidental nature of a constraint against English morphemes ending in two voiced obstruents cannot be maintained.

What Kahn (1976) has actually pointed out is one way in which to determine that a constraint is an MSC rather than an SSC. Specifically, if there are no occurrences of a particular type of monomorphemic monosyllable (e.g. those ending in two voiced obstruents) but there are occurrences of such monosyllables that are bimorphemic, then the restriction being dealt with is one pertaining to morphemes and not to syllables.

Another way of determining whether a constraint is an MSC rather than an SSC is by examining polysyllabic monomorphemic words. For example, if a constraint is posited between two segments based on monosyllabic monomorphemic words (where the two segments would be members of the same syllable), and that constraint is also relevant for the same segments in polysyllabic monomorphemic words (in which the two segments would be members of different syllable) then that constraint is an MSC rather than an SSC. If, on the other hand, the constraint is not relevant for the same segments in polysyllabic monomorphemes (i.e. the constraint holds between two sounds in the same syllable but not when they are in different syllables), then the constraint reflects an SSC. An example that can help elucidate this comes from the MSC's that hold among root consonants in Arabic which Hooper (1975) seems to be unaware of. Arabic has restrictions on what consonants can cooccur in a root. Greenberg (1950) notes restrictions such as the following: two post-velars cannot occur in the same root, the first two root consonants cannot be homorganic obstruents. The syllable plays no role in such constraints. This is because root consonants in a word can all be in one syllable, or in two syllables, or even in three different syllables. Thus the constraints on root consonants in Arabic are not syllable-sensitive, rather they reflect MSC's. Examples like Arabic clearly show that, despite Hooper's (1975) contention, MSC's cannot always be reduced to SSC's.

Another way of determining whether a constraint is an MSC or an SSC concerns instances where position within the word is the main factor in whether or not two sounds cooccur. If the two sounds can cooccur in a word only if they are heterosyllabic, then the restriction that prevents them from occurring within the same syllable of the word is a reflection of an SSC. A good example of such an instance is the restriction in English on *tl* and *dl*. In English no morphemes begin with *tl*. However, sequences of *tl* do occur when they are heterosyllabic as in words like *atlas* and *Atlantic* (or even in morpheme-final position as in the name *Aristotle* where the stress pattern–primary stress on the first syllable–indicates that the final *l* is not underlyingly syllabic). The restriction on English *tl* (as well as on *dl*) thus reflects an SSC that prohibits such sequences in syllable onsets.

Based on our discussion so far, it can be concluded that languages can have both MSC's and SSC's and that there are means for determining whether a given constraint is an MSC or an SSC. Besides applying over different domains (morpheme vs. syllable), MSC's and SSC's differ from one another in another important way. MSC's pertain to underlying representation whereas SSC's come into play at the point in the derivation in which syllable formation rules apply. We assume here a rule-based account of syllabification along the lines of Steriade (1982). Because MSC's and SSC's are relevant at different points in the derivation, they may assume different representations for the same phoneme. For example, if the feature (or node) [coronal] is unspecified in underlying representation, it would be predicted that MSC's could not refer to the feature (or node) [coronal] but SSC's could refer to [coronal] as long as specification of [coronal] occurs before the application of the syllable formation rules. In the following section we consider the case of English where it is shown that MSC's treat some coronals such as /t/ as lacking the Place Node. As noted above, English SSC's do not necessarily treat /t/ as lacking it.

In summary, in this section we have shown that despite Hooper's (1975) contention, MSC's are not always reducible to SSC's. Moreover, we have pointed out various ways to determine whether a restriction holding between two segments reflects an MSC or an SSC. Finally, it has been suggested that MSC's reflect the nature of underlying representations while SSC's reflect the nature of representations at the point in the derivation where syllable formation rules apply.

The Underspecification of Coronals in English

Most of the recent work examining what is underspecified in English has not used MSC's as a criteria–with the notable exception of Clements (1988). As mentioned earlier, Avery & Rice (1989) base what is specified on the nature of the phonemic inventory. They contend that for English the anterior coronals /t/ and /n/ are underspecified for place of articulation since these phonemes do not contrast with nonanterior coronals. On the other hand, /s/ would be specified for place of articulation (i.e. it has the Coronal Node) since it does contrast with the nonanterior coronal /š/. Avery & Rice find support for their view of English underspecification from phonological rules that seem to treat /t/ and /n/ as if they do not possess the Place Node. The specific rules that Avery & Rice mention are a rule that turns (syllable-final) /t/ into [ʔ] in such words as *button* and *cotton* and a rule that optionally assimilates word-final /n/ to the place of articulation of the following word-initial consonant. However, it can be maintained that the former rule really does not show that /t/ is underspecified for place of articulation features. This is because the rule affects syllable-final /t/, which means that rules of syllabification (and resyllabification) have already applied before the /t/-to-[ʔ] rule. Since syllabification has already applied then SSC's have also already taken effect. Because English has an SSC that specifically refers to the Coronal Node of /t/ and /d/ (i.e. the one prohibiting syllable-initial /tl/ and /dl/ sequences), the Coronal Node must be present at the time the /t/-to-[ʔ] rule applies. Consequently, this rule should be interpreted as simple delinking of the Supralaryngeal Node of /t/, and it would not constitute evidence that /t/ is unspecified for the Place Node in underlying representation. Moreover, Avery & Rice's contention that specification of the Coronal Node can be determined by the presence of an anterior/nonanterior contrast is called into question by Paradis & Prunet's (1989b) work on Fula. They argue that in Fula, a language that contrasts both anterior and nonanterior coronals, the anterior coronals lack the Place Node altogether since they are the only consonants that are transparent to a process of vowel spreading. They conclude that all ([+anterior]) coronals lack the Place Node as a principle of grammar

In this Section, we consider MSC's as a criteria for determining what features (or Nodes) are unspecified in underlying representation. If we use MSC's as criteria for determing underspecification, rather than phonological rules or the nature of phonological inventories, it becomes more readily apparent what features (and Nodes) are present in underlying representation. This is because MSC's hold on underlying representations prior to phonological or morphological processes. The specific question to be addressed in this section is whether or not MSC's treat coronals as "special". By special, we mean that coronals (or, at least, some coronals) are treated as if their Place Node is absent in underlying representation. We show that despite claims to the contrary by Mester & Ito (1989) MSC's can treat coronals as special. Specifically, we show that English MSC's treat coronal stops as special.

Recently, it has been argued explicitly by Mester & Ito (1989) (and implicitly by Clements 1988) that MSC's do not treat coronals as special even in languages where there is no contrast between anterior and nonanterior coronals. For example, in Classical Arabic (Greenberg 1950 and McCarthy 1988), which has no contrast between anterior coronal stops and nonanterior coronal stops, there is a constraint that rules out homorganic consonants (or, more accurately, obstruents made with the same articulator) from occurring in the same root morpheme. This MSC holds for all places of articulation including coronal. Coronal consonants are not ignored by this constraint. Because Arabic coronal obstruents are subject to an MSC that prevents them from occurring with other coronal obstruents in the same root, the Coronal Node must be present in the Underlying Representation. Consequently, MSC's like that in Arabic involving place of articulation argue against a specific place of articulation (like coronal) being completely unspecified in underlying representation. This is pointed out by Mester & Ito (1989) who note that with such Morpheme Structure Conditions, "... no special status is accorded to the unmarked place, whatever it may be."

If special status were accorded to an unmarked place (i.e. coronal), it would be expected that MSC's which pertain to homorganic consonants would not hold for consonants of the unmarked place. So, for example, if in some hypothetical language it is posited that a coronal consonant such as /t/ has no Place Node and if that language possesses MSC's of the sort found in Arabic, then it would be predicted that, in general, morphemes would not contain homorganic consonants; however, this prediction would not hold for /t/ since it would not be represented with the Place Node. The existence of such a hypothetical language would lead to the conclusion that because of the type of MSC's found in a language like Arabic, it is a parameterized option whether or not a language can have the Place Node completely absent in underlying representation. We now consider a type of MSC found in English that shows that English is our hypothetical language.

Fudge (1969), Clements & Keyser (1983), and Davis (1984) have all observed a number of constraints that restricts the type of consonant that flanks both sides of a vowel in sCVC sequences. One of the strongest of these constraints is that in sCVC monosyllables, the same noncoronal consonant cannot flank both sides of the vowel. Hence, there are no English words like *spap*, *spep*, or *skik*. On the other hand, monosyllables with the same coronal flanking both sides of the vowel in a sCVC word do occur (eg, state, stout, stoat).

Fudge (1969), Clements & Keyser (1983), and Davis (1984) express the sCVC constraint as an SSC. However, it will be shown that this constraint is not a condition on syllables, but is rather a reflection of an MSC. Afterwards, it will be shown that the condition pertains to homorganic consonants flanking both sides of the vowel rather than just to identical consonants.

If, the constraint on sCVC sequences were a reflection of an SSC, one would expect to find English words containing the sequence sCVCV since the postvocalic C would not be part of the initial syllable. So, for example, one might expect that there would be words like *spapoon* or *skikanda* in which the first postvocalic consonant is not part of the initial syllable. If, as is argued here, the constraint against sCVC sequences is a reflection of an MSC then possible monomorphemic forms like *spapoon* or *skikanda* would never occur (or, at least be extremely rare). In order to determine whether such English monomorphemes occur, a search was conducted on a computerized lexicon containing nearly 20,000 words from Webster's Pocket Dictionary. The only word in this lexicon in which the sequence sCVC was found (where the C's are identical noncoronal consonants) was the word *dyspepsia* where the sequence *spep* occurs. However, the sequence *spep* in this word spans a morpheme boundary since the initial /s/ is part of the morpheme *dys* which also occurs in words like *dysfunction* and *dystrophy*. Thus no monomorphemic sCVC sequences were found in which the two C's were identical noncoronal consonants. Consequently, the constraint on sCVC sequences in English seems truly to be a reflection of an MSC. Moreover, this MSC is indeed restricted to noncoronals because in addition to the monosyllabic morphemes mentioned above, like *state stout* and *stoat* where the coronal /t/ flanks both sides of the vowel, there are monomorphemic stVt sequences in such words as *astute*, *statistics*, *status*, *stutter* and *substitute*.

For sake of completeness, we note that morphemes having the sequence sCVC where the two C's are different noncoronals are common. A search through the 20,000 word lexicon gives us such words as *speak*, *skip*, *spaghetti*, *scaffold*, *scuba*, *eskimo*, and *episcopal*. Also, morphemes having the sequence sCVC where the first C is coronal and the second C noncoronal are common. Such forms include *stake*, *stop*, *stable* and *stagger*. Thus, with the data discussed so far, it can be concluded that the constraint on sCVC sequences is a reflection of an MSC holding between identical noncoronal consonants.

The MSC pertaining to sCVC-sequences on further investigation turns out to be a more general constraint in that it rules out morphemes where the two C's are homorganic, not merely identical. That is, there are virtually no monomorphemic forms in English that have the sequence sCVC where the two C's are either both labial or both velar. The only word in the 20,000 word lexicon that was found to violate this constraint is the word "skunk" (on the assumption that English has underlying velar nasals).[1] That this constraint really does involve identical place of articulation is made evident when we consider the situation where the two C's in an sCVC sequence are not homorganic. A search through the 20,000 word computerized lexicon revealed that no constraint whatsoever held when the two C's were made at different locations in the vocal tract. For example, the sequence skV was followed by a labial consonant in 58 entries (*skip*, *scuba*), an alveolar consonant in 151 entries (*skit*, *skate*), and a palato-alveolar consonant in 25 entries (*scotch*, *sketch*). The fact that there are virtually no words with a velar consonant following an skV sequence is of interest. Moreover, the sequence spV was followed by a velar consonant in 56 entries (*spike*, *spook*), an alveolar consonant in 196 entries (*spit*, *speed*), and a palato-alveolar consonant in 20 entries (*speech*, *special*); there are virtually no words where a labial consonant followed an spV sequence.

Finally, while the constraint on sCVC sequences holds for homorganic noncoronals it clearly does not hold for /t/. Many morphemes have the sequence stVC where the postvocalic C is a coronal. There were over one hundred entries in which the coronal was an obstruent and over two hundred entries in which the coronal was a sonorant. Typical examples include *stud*, *study*, *astound*, *stadium*, *stash*, *stitch*, and *stone*. Furthermore, there were over one hundred entries in which the postvocalic C was a labial (*stable*, *stop*) and over one hundred entries in which it was a velar (*stock*, *plastic*). Thus, English has an MSC that prevents homorganic noncoronals from flanking both sides of the vowel in sCVC sequences.

Consequently, it is concluded that, contrary to what Mester & Ito (1989) contend, MSC's can treat coronal consonants as special. The English MSC discussed in this paper can only be understood if /t/ in English lacks the Place Node but labial and dorsal consonants do not.[2] For it is only /t/ that is not subject to the MSC that prevents homorganic consonants from flanking both sides of the vowel in sCVC sequences.[3]

---

[1] There are a handful of other words that come to mind that violate the constraint but which do not appear in the computerized lexicon. These include *spam*, *spumoni* *spoof* and *spiffy*. It may be that the MSC preventing two homorganic noncoronal consonants from occurring in sCVC sequences is "tighter" if the two consonants are both oral stops.

[2] It is also possible to conclude that English /t/ does not lack the Place Node rather it lacks the articulator node Coronal. While I am unaware of evidence from English MSC's that would help determine this, I am assuming that it is the Place Node that is lacking. This is the case for other languages such as Fula where Paradis & Prunet (1989b) show that for ([+anterior]) coronals it must be the Place Node that is lacking (and not just the articulator node) in order for such consonats to be completely transparent to vowel spreading.

[3] It is interesting to note that while English has an MSC on sCVC sequences there appear to be no systematic constraints on CVC sequences. Such monosyllables as *pipe*, *kick*, *tight*, *pub*, *cog*, and *toad* with homorganic consonants flanking both sides of the vowel occur in CVC sequences. We repress the temptation to speculate on why the MSC only holds for sCVC sequences, but we do note that the constraint is not idiosyncratic. MSC's on the homorganicity of other consonants are found in other languages such as Arabic

Although we have so far argued that /t/ lacks the Place Node in English we have yet to focus on other coronal consonants. It is briefly noted here that the evidence from other MSC's in English is not incompatible with a view that coronal sonorants (/n/, /l/, and /r/) lack the Place Node in underlying representation, whereas coronal stridents do not.

The MSC evidence relevant for coronal sonorants is inconclusive regarding whether these sounds lack the Place Node. Consider, first, the coronal nasal /n/. English has an MSC that prohibits sNVN sequences (where N=any nasal). The coronal nasal is not exceptional to this constraint. There are no sequences like *snan* in English monomorphemes. This MSC, though, only implies that all nasal consonants in English must have the feature [+nasal] in underlying representation. But this does not at all imply the presence of the Place Node for /n/ (at least under a view of feature geometry in which [nasal] is located immediately under the Root Node or the Supralaryngeal Node). As for /l/ and /r/, English has an MSC that prohibits identical liquids from occurring in CLVL sequences (where L=liquid). Thus potential sequences like *plil* or *bror* do not occur in English monomorphemes. The only exception is *slalom* (although *flail* would also be exceptional if it is pronounced with a single vowel). This MSC, though, only implies the presence of the feature [lateral] in underlying representation for the phonemes /l/ and /r/. It does not, however, imply the presence of the Coronal Node.[4] Thus English MSC's relating to coronal sonorants are not incompatible with a view that these sonorants lack the Place Node.

The evidence that coronal stridents require the Place Node in underlying representation comes from an MSC that has been discussed by Clements (1988). He notes that English roots do not contain adjacent coronal stridents that are both [+continuant].[5] This MSC assumes the presence of the Coronal Node (and, consequently, the Place Node) in the underlying representation of stridents.

In conclusion, based on the evidence from the English MSC's discussed in this section, /t/ (and assumingly /d/) lack the Place Node in underlying representation but coronal stridents do not. The MSC evidence is inconclusive concerning the lack of the Place Node in coronal sonorants.[6] Nonetheless, we have found in this paper that MSC's can, and do, treat coronals as special. This finding shows that the contention of Mester & Ito (1989) that "...homorganicity restrictions hold for ALL places of articulation, and no special status is

---

(McCarthy 1988) Javanese (Mester 1986) and Cambodian (Yip 1989).

[4] Levin (1988) has argued that the feature [lateral] is dominated by the Coronal Node, so that the presence of the feature [lateral] implies the presence of the Coronal Node. However, here we follow the position of Shaw (1988) in which it is argued that [lateral] cannot be dominated by the Coronal Node but is located higher up in the feature geometry tree.

[5] The MSC holds on a sequence of two strident fricatives, a sequence of a strident affricate and a strident fricative, but not on a sequence of a strident fricative followed by a strident affricate (e.g. /šs/ and /čs/ do not occur, but /sč/ does occur as in *eschew*). This can be taken as evidence that affricates have the feature values [-cont], [+cont] and that these features are sequentially ordered.

[6] The interdental phonemes /θ/ and /ð/ are not dealt with here because of their low frequency of occurrence.

accorded to the unmarked place..." cannot be maintained.

Moreover, our finding argues against the view of underspecification advanced by Clements (1988) in which the Place Node is required to be present in underlying representation. The English MSC forbidding homorganic noncoronals in sCVC sequences is best understood only if /t/ is represented without the Place Node in underlying representation.

However, our finding for English is basically compatible with either a contrastive theory of underspecification (e.g. Avery & Rice 1989) or more radical theories of underspecification (e.g. Archangeli 1988). It is compatible with a contrastive theory of underspecification because the only MSC that seems to require the presence of the Place Node for coronal consonants is the MSC noted by Clements (and discussed above) which prohibits two adjacent coronal stridents that are both [+continuant]. Coronal stridents are the only coronal consonants in English that are contrastive for the feature [anterior] (ignoring the problem of how English affricates–which are coronal stridents– should be represented). So the Coronal Node would at least be posited for these consonants anyway under a contrastive theory of underspecifcation. Our finding is also compatible with more radical versions underspecification since some of the English coronal consonants must lack the Place Node.

While, as mentioned, our specific finding for English is basically compatible with either a contrastive theory of underspecification or more radical theories, the MSC's on homorganicity of root consonants in Arabic discussed earlier seem incompatible with both. This is because in Arabic there is no contrast between the voiceless anterior coronal stop /t/ and a corresponding nonanterior coronal. So both contrastive and radical theories of underspecification would apparently posit that the Arabic /t/ should be represented without the Coronal Node in underlying representation. But because Arabic /t/ is subject to an MSC that prevents it from occurring with other coronal obstruents in the same root, the Coronal Node must be present in the underlying representation of /t/. Consequently, the different realizations of MSC's that are found in languages like English and Arabic provide support for a theory of underspecification in which the presence of the Place Node for coronals is a parameterized option. At least some English coronal consonants lack the Place Node whereas Arabic coronal consonants do not.

References

Anderson, S. (1982). Where's morphology? *Linguistic Inquiry*, **13**, 571-612.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology*, **5**, 183-297.

Avery, P. & Rice, K. (1989). Segment structure and coronal underspecification. To appear in *Phonology*, **6.2**.

Clements. G. (1988). Towards a substantive theory of feature specification. *North Eastern Linguistic Society*, **18**, 79-93.

Clements, G. & Keyser, S. (1983). *CV phonology*. Cambridge, MA: MIT Press.

Davis, S. (1984). Some implications of onset-coda constraints for syllable phonology. *Chicago Linguistic Society*, **20**, 46-51.

Davis, S. (1988). Manner of articulation and feature geometry: A phonological perspective. *Research on speech perception progress report no. 14*. Bloomington IN: Speech Research Laboratory, Psychology Department, Indiana University, pp. 179-207.

Fudge, E. (1969). Syllables. *Journal of Linguistics*, **5**, 253-287.

Greenberg, J. (1950). The patterning of root morphemes in Semitic. *Word*, **6**, 162-181.

Hooper, J. (1975). The archi-segment in natural generative phonology. *Language*, **51**, 536-560.

Janda, R. (1983). Morphemes aren't something that grows on trees: Morphology as more the phonology than the syntax of words. *Chicago Linguistic Society*, **19** (Parasession), 79-95.

Kahn, D. (1976). Syllable-based generalizations in English phonology. Doctoral dissertation, MIT, Cambridge, MA. Distributed by the Indiana University Linguistics Club, Bloomington, Indiana.

Levin, J. (1988). A place for lateral in the feature geometry. Unpublished manuscript. Department of Linguistics, University of Texas, Austin, Texas.

McCarthy, J. (1988). Feature geometry and dependency. *Phonetica*, **43**.

Mester, R. (1986). Studies in tier structure. Doctoral dissertation, University of Massachusetts, Amherst, Massachusetts.

Mester, R. & Ito, J. (1989). Feature predictability and underspecification palatal prosody in Japanese mimetics. *Language*, **65**, 258-293.

Paradis, C. & Prunet, J.-F. (1989a). Markedness and coronal structure. *North Eastern Linguistic Society*, **19**.

Paradis, C. & Prunet, J.-F. (1989b). On coronal transparency. To appear in *Phonology*, **6.2**.

Sagey, E. (1986). The representations of features and relations in nonlinear phonology. Doctoral dissertation, MIT, Cambridge, Massachusetts.

Shaw, P. (1988). The locus of lateral in feature geometry. Paper presented at the Annual Meeting of the Linguistic Society of America, Dec. 27-30, New Orleans, Louisiana.

Steriade, D. (1982). Greek prosodies and the nature of syllabification. Doctoral dissertation, MIT, Cambridge, Massachusetts.

Yip, M. (1989). Feature geometry and co-occurrence restrictions. To appear in *Phonology*, **6.2**.

II. SHORT REPORTS AND WORK-IN-PROGRESS

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Position of the Maximum Amplitude as a Perceptual Stress Cue in English: Work in Progress[1]

Dawn M. Behne

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported, in part, by NIH Training Grant No.NS07134-11 to Indiana University in Bloomington, IN. The author would like to thank W. Charles Read of the University of Wisconsin-Madison and the rest of the dissertation committee for discussion leading to this investigation.

Abstract

Early discussions of the acoustic correlates of stress suggested that stressed syllables are produced with greater amplitude than unstressed syllables. More recently, increased fundamental frequency (F0) change and duration have been demonstrated to be more reliable cues of stress than increased amplitude. Nevertheless, none of these acoustic parameters has been consistently connected with stress, and their relative importance as cues for stress appears to vary across languages. Behne (1989) found the early position of the maximum amplitude to be associated with stress in production. The present study is an initial investigation of early maximum amplitude as a perceptual cue for stress in English and begins to address four interrelated points:

- Is an early maximum amplitude able to serve as a perceptual cue for stress?
- If so, does the relative maximum amplitude positions in two successive syllables affect stress perception?
- How great must the maximum amplitude position difference be in two successive syllables for a stress difference to be perceived?
- Does the maximum amplitude position affect stress perception of different vowels in the same manner?

To date, the perception results from English support the production research showing an early maximum amplitude to be associated with stress. The results suggest that early discussions of amplitude as an important acoustic correlate of stress may not have been far off, but that the position - rather than the relative level - of the maximum amplitude is the pertinent parameter.

Position of the Maximum Amplitude as a Perceptual Stress Cue in English: Work in Progress

Stress refers to both the production and perception of prosodic salience. The production of stress is not clearly attributable to a unique speech mechanism, but it has been closely linked to a speaker using more muscular effort and energy than when there is no stress (Armstrong 1932; Coustenoble & Armstrong 1934; Ladefoged 1967; Lehiste 1970). The resulting physiological state (e.g. increased subglottal pressure, vocal fold tension, and pulmonary effort) gives rise to acoustic phenomena which are perceived as having greater prominence (Ladefoged 1967; Lehiste 1970).

The smallest unit of speech which can be more stressed than another is the syllable. Inherent to a syllable are a basic F0, duration and amplitude. When syllables occur in succession in speech, these acoustic parameters are adjusted in such a way that some syllables become more salient than others. These acoustic properties appear to be carried primarily by the nucleus of the syllable (Fry 1955; Oller 1973).[1]

Increased F0 Change, Duration and Maximum Amplitude as Stress Cues

Early discussions of the acoustic correlates of stress suggested that stressed syllables are produced with greater intensity than unstressed syllables (Sweet 1890; Bloomfield 1933; Jones 1960).[2] Although it was also suggested that stressed syllables are longer than unstressed syllables (Armstrong 1932; Paramenter & Trevino 1935) and that pitch change was correlated with stress (Passy 1907; Stetson 1928; Sweet 1890), these were usually considered only secondary cues to stress.

More recent investigations have demonstrated that intensity is not likely the important acoustic correlate of stress it was originally believed to be, and that F0 change and duration are stronger cues to stress than intensity. Fry (1955) synthesized English noun/verb pairs such as "OBject/obJECT", systematically varying the vowel duration and intensity of both syllables. Listeners perceived long, high intensity syllables as stressed, and short, low intensity syllables as unstressed. With duration and intensity varying separately, vowel duration provided listeners with a better cue to stress than intensity. Fry (1958) confirmed these results and demonstrated that the presence of an F0 change was an even stronger perceptual cue for stress than duration. A series of studies summarized in Bolinger (1958) lead to the conclusion that F0 change was the primary stress cue, and although duration covaried with F0, intensity was not relevant as a cue to stress. Using synthetic nonsense syllables Morton

---

[1] Although the smallest possible syllable has only the nucleus, which also seems to carry the acoustic parameters associated with stress, strickly speaking the syllable should still be considered the smallest domain of stress since consonants surrounding the nucleus can also influence acoustic properties of the syllable (e.g. House & Fairbanks 1953; Delattre 1963)

[2] This gave rise to the term "accent of intensity"

& Jassem (1965) also systematically varied F0, duration and intensity. As in Bolinger (1958) and Fry (1958), F0 change was found to be the strongest perceptual cue, and although more intense, longer syllables were also likely to be perceived as stressed, the relative importance of duration and intensity was not evident.

Acoustic investigation of natural speech has provided support for the results of perceptual research. Lieberman (1960) analyzed noun/verb pairs like those used in Fry (1955, 1958). Although he found F0 to be most strongly correlated with stress, intensity was found to be more closely associated with stress than vowel duration.

In these studies, F0 was consistently found to provide the best cue for stress among the parameters investigated. However, Cutler & Darwin (1981) have shown that stress is still perceived if F0 is held constant. That is, although syllables having a F0 change are more likely to be perceived as stressed, F0 change does not need to be present for listeners to perceive stress.

In Behne (1989), these acoustic parameters were used as the basis for an investigation of stress in a production task comparing English and French. As in English, previous research for French has suggested that F0 change is most closely correlated with stress; although in English increased vowel duration appears to be a stronger cue of stress than increased maximum amplitude, they seem to be equally important cues of stress in French (e.g. Rigault 1962).[3] The results generally confirmed these parameters as cues of stress for English, but not for French. For English, the results show clear evidence of stress from increased F0 change and vowel duration, but not from an increased maximum amplitude. In French there was little evidence of stress from increased F0 change, vowel duration or increased maximum amplitude. The results were interpreted as evidence that the traditional acoustic parameters are not sufficient for cuing stress.

Research into the acoustic nature of stress has not clearly shown that increased F0 change, duration and maximum amplitude are the only acoustic parameters relevant to stress; the results for English have not been consistent and the importance of these traditional acoustic correlates of stress appears to vary across languages.

Position of the Maximum Amplitude as a Stress Cue in Production

Behne (1989) investigated the shapes of the F0 and amplitude contours for stressed and unstressed syllables. The F0 and amplitude contours were measured extensively in order to develop models of the contours for stressed and unstressed syllables. By examining the proximity and relative positions of the maximum and minimum F0 and amplitude, models were used to explore the association of the contours' shapes with stress for English and

---

[3] For a more indepth review of the literature on acoustic correlates of stress in French see Chapter III in Behne (1989).

French. Investigation of F0 and amplitude contours revealed acoustic correlates of stress beyond those which have traditionally been studied. In these contours the position of the maximum amplitude was consistently associated with phrase-final and focal stress in both English and French. In both languages the maximum amplitude occurred earlier in stressed vowels than in unstressed vowels. The results are summarized in Table 1.[4]

Insert Table 1 about here

Based on these results, the present study was designed as a preliminary investigation of the maximum amplitude position as a perceptual cue for stress. Although plans are being made to extend this line of research to French, the present study focuses only on English. Four interrelated points are addressed:

- Does the relative maximum amplitude positions in two successive syllables affect stress perception?
- How great must the maximum amplitude position difference be in two successive syllables for a stress difference to be perceived?
- Does the maximum amplitude position affect stress perception of different vowels in the same manner?
- Is an early maximum amplitude able to serve as a perceptual cue for stress?

Method

Stimuli. The vowels /i, e, u, o/ were produced using the Klatt Synthesizer which allows F0, formant frequencies and amplitude to be manipulated every 5 msec. All four vowels had a duration of 345 msec. All vowels also had the F0 contour presented in Figure 1, with the F0 rising from 0 Hz at the start of the vowel to 98 Hz at 5 msec, and 100 Hz at 50 msec staying at 100 Hz until 300 msec, then falling to 98 Hz at 340 msec and 0 Hz at 345 msec. The formant frequencies for the four vowels are presented in Figure 2.

Insert Figures 1 and 2 about here

---

[4] In addition, in English stress was associated with a falling amplitude contour, and in French, stress was associated with a falling F0 contour.

Table 1

The mean relative position of the maximum amplitude for English from Behne (1989) presented in terms of the percent into the vowel.

	English	French
Stressed	33.4%	45.2%
Unstressed	42.3%	49.6%

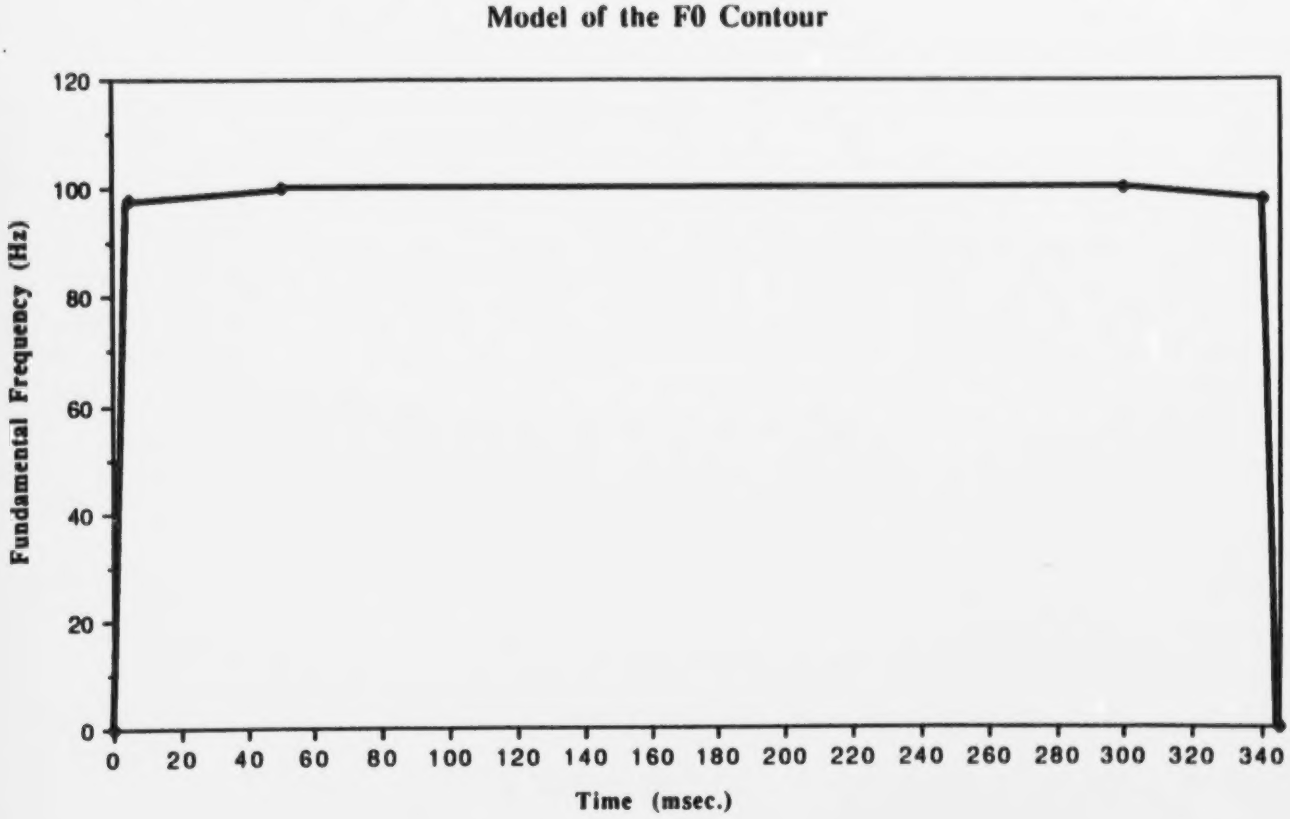


Figure 1. The fundamental frequency contour used for all of the synthesized vowels.

```
                                                                    VOWEL POSITION (msec)

                    0          50         100     120   130      150        200        250   260    270    300         345
-------------------------------------------------------------------------------------------------------------------------
/i/    F1          310 ----- 310 .................................................... 290 -------------------------- 290
       F2         2020 ---- 2020 ................................................... 2070 -------------------------- 2070
       F3         2960 ---- 2960 ................................................... 2980 -------------------------- 2980
       F4         3400 ---------------------------------------------------------------------------------------------- 3400

/e/    F1          480 ------------------------------ 480........................................ 400--------------- 400
       F2         1720 ----------------------------- 1720....................................... 2020--------------- 2020
       F3         2520 ----------------------------- 2520....................................... 2600--------------- 2600
       F4         3500 ---------------------------------------------------------------------------------------------- 3500

/u/    F1          350 ------------------------ 250 ...................................... 320---------------------- 320
       F2         1250 --------------- 1250 .............................................. 900---------------------- 900
       F3         2200 ---------------------------------------------------------------------------------------------- 2200
       F4         3500 ---------------------------------------------------------------------------------------------- 3500

/o/    F1          540 ---------------- 540 ......................................... 450 -------------------------- 450
       F2         1100 --------------- 1100 ......................................... 900 -------------------------- 900
       F3         2300 ---------------------------------------------------------------------------------------------- 2300
       F4         3500 ---------------------------------------------------------------------------------------------- 3500
```

Figure 2. Formant variations of the synthesized vowels /i, e, u, o/.

For each of the four synthesized vowels, a set of 44 vowels was generated which differred only in their amplitude contours, as is represented in Figure 3. The amplitude contours of all 176 vowels (4 vowel types x 44 vowels in each set) had a 45 msec onset (0-45 msec) and offset (300-345 msec), during which time the amplitude linearly increased and decreased respectively between 0 and 50 dB. All vowels also had a maximum amplitude of 70 dB, with linear interpolation from the onset to the maximum and from the maximum to the offset. The only parameter of the amplitude contour which varied was the position of the maximum amplitude, that is the distance of the maximum amplitude from the beginning of that vowel. Within each vowel set, the position of the maximum amplitude shifted from 50 msec to 265 msec into the vowel in 44 steps of 5 msec, as is represented in Figure 3. For example, in the first vowel within each of the four vowel sets, the amplitude rose from 0 dB at 0 msec to 50 dB at 45 msec, then rose to the maximum of 70 dB at 50 msec and gradually dropped to 50 dB at 300 msec, and then to 0 dB at 345 msec. In the second vowel of each set, the onset and offset remained the same but the amplitude rose from 50 dB at 45 msec to 70 dB at 55 msec, then dropped to 50 dB at 300 msec. In this manner, the position of the maximum amplitude was shifted 5 msec later into the vowel successively for the 44 vowels in each set.

Insert Figure 3 about here

The 176 synthesized vowels were used to form vowel pairs. Every pair was composed of two vowels which were qualitatively alike and separated by 50 msec of silence. Vowel pairs were formed based on two factors of the maximum amplitude position: (1) amount different, that is the amount of difference between the maximum amplitude positions of the first and second vowel in a pair; and (2) direction different, that is whether the maximum amplitude position in the first vowel of the the pair is relatively earlier (A<B) or relatively later (A>B) than that of the second vowel of the pair.

Within each of the four vowel sets, every vowel was coupled with each vowel which had a maximum amplitude in a position 10, 15, 20, 25, 30, 35, 40, 45 and 50 msec different from itself, introducing nine experimental levels of amount different. For each of the nine levels, the two vowels of each pair were arranged in two ways to introduce the second factor, direction different; either the first vowel had an earlier maximum amplitude than the second vowel, or the first vowel had a later maximum amplitude than the second vowel.

In addition, control vowel pairs were developed by pairing each of the 176 synthesized vowels with a copy of itself. The control vowel pairs were intermixed with 64 filler vowel pairs, 16 pairs for each of the 4 vowels. In the 16 pairs for each vowel, the first vowel in each pair had a maximum amplitude positioned 105, 110, 115, 120, 125, 130, 135 or 140 msec from the beginning of the vowel, and the second vowel had a maximum amplitude positioned either 50 msec before or after that of the first vowel of the pair.

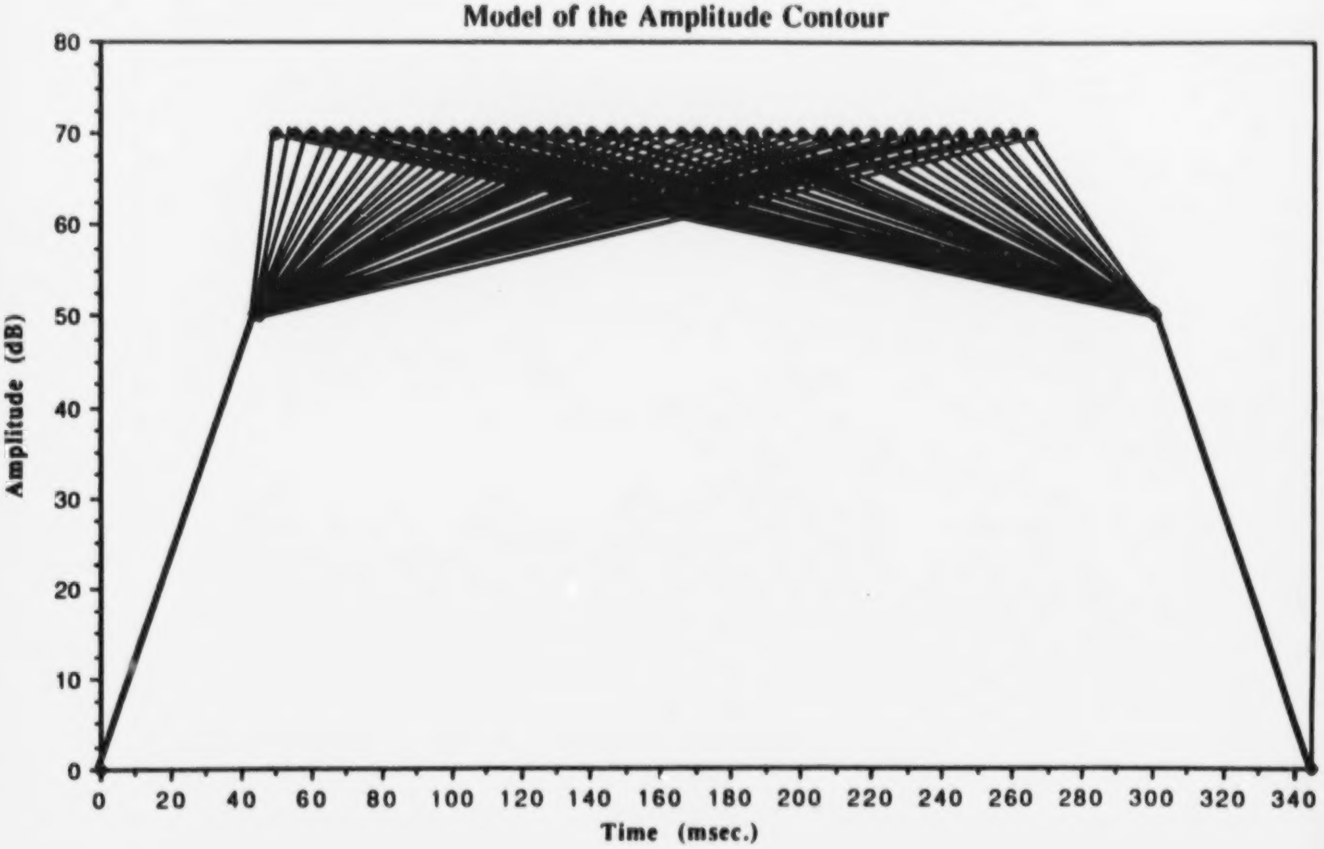


Figure 3. Model of the amplitude contour for the four sets of vowels, with the onset and offset remaining constant and the position of the maximum amplitude shifting within each vowel set in 44 steps from 50 msec. to 265 msec. into the vowel.

Subjects. One hundred introductory psychology students at Indiana University participated in the study. All of the participants were monolingual, native speakers of American English.

Procedure. For each of the ten levels (i.e. the control and the nine experimental levels) of amount different, A<B and A>B pairs of the four vowels were randomized and presented to ten subjects.

Each subject was seated in a booth which had a headset and a two-button response box. Subjects were instructed that pairs of vowels would be presented over the headset and that they were to decide as quickly as possible which vowel in each pair was strongest. If the first vowel was strongest, the left button would be pushed; if the second vowel was strongest, the right button would be pushed. Subjects were told to make a response for all vowel pairs. Before starting, subjects heard and responded to five practice vowel pairs. Subjects had up to 5 seconds to respond to a vowel pair before a light signaled that the next vowel pair was about to be presented. Subjects' responses were automatically recorded.

Results and Conclusions

The expectation that, of the two vowels in each pair, the member with the earlier maximum amplitude would be perceived as stressed was used as the baseline for tabulating subjects' responses. In A<B pairs, the maximum amplitude is earlier in A than in B, and in A>B pairs, the maximum amplitude is earlier in B than in A; the number of A responses for A<B pairs and the number of B responses for A>B pairs were tabulated.

To date, the data have not been fully analyzed; however the general tendencies of the results are discussed here from two perspectives. First, within each of the ten levels of amount different, separate omnibus analyses of variance were conducted for the two levels of directions different (A<B) and (A>B).[5] In each analysis of variance, there were two factors: (1) maximum amplitude position, using the member of the vowel pair with the earlier maximum amplitude as a reference point, and (2) vowel type (i.e. /i, e, u, o/). The F-values are presented in Table 2.

Second, the data will be considered in terms of the results in Behne (1989). As is partially shown in Table 1, Behne (1989) found the position of the maximum amplitude in English to be an average of 33.4%, but as early as 29.9%, into a stressed vowel, and an average of 42.3%, but as late as 47.7%, into an unstressed vowel. In terms of the 345 msec vowels in the present study, vowels with maximum amplitudes positioned from approximately 105 msec (30.0%) to 115 msec (32.9%) are expected to be perceived as stressed compared to those with maximum amplitudes positioned from 150 msec (42.9%) to 165 msec (47.1%), a difference in position of at least 35 msec. For the purposes of this progress report, the range of maximum amplitude positions from 105 msec to 165 msec, and the amount different from 35 msec to 50 msec will be referred to as the delimited data. Lines of best fit for these data are presented in Figures 5 through 12.

---

[5] The significance level of 0.05 is accepted as a standard for this investigation. Although not accepted as significant, F-values at $0.25<p<0.05$ are noted.

The Control Condition

Two points should be noted concerning the data from the control group. First, as would be expected, the maximum amplitude position is not a significant source of variance for the control group [F(43,1584)=0.54; n.s.]. This finding indicates that stress perception of identical vowels does not generally vary systematically across the different maximum amplitude positions. Second, vowel type is a significant source of variance for the control group [F(3,1584)=8.95; n.s.]. The lines of best fit in Figure 4 suggest that subjects perceive the first member of /o/ vowel pairs as slightly more stressed than the second, and the second member of /i/ vowel pairs as more stressed than the first, with closer to chance responses for /u/ and /e/. Although the interaction of maximum amplitude position and vowel type is not significant [F(129,15840=0.9008; n.s.]., these tendencies appear to become stronger as the maximum amplitude position becomes later. This point will be-explored further with more focused statistical analysis of the means and be discussed in a later report.

Insert Figure 4 about here

Does the relative position of the maximum amplitude in two successive syllables affect stress perception?

The analyses performed to date do not directly address a difference between A<B and A>B vowel pairs; however, as is shown in Table 2, the maximum amplitude position is a significant source of variance more frequently across the levels of amount different for A>B vowel pairs than for A<B vowel pairs. A comparison of Figures 5-8 with Figures 9-12 illustrate a strong tendency within the delimited data for A<B vowel pairs to receive fewer than chance first vowel responses, but for A>B vowel pairs to receive greater than chance second vowel responses. The data suggest that an early amplitude peak is more effective as a cue for stress in the second of two successive syllables. This point will be explored further with more focused statistical analyses and in a later report conclusions will be discussed in terms of other declination effects for acoustic correlates of stress.

Insert Table 2 about here

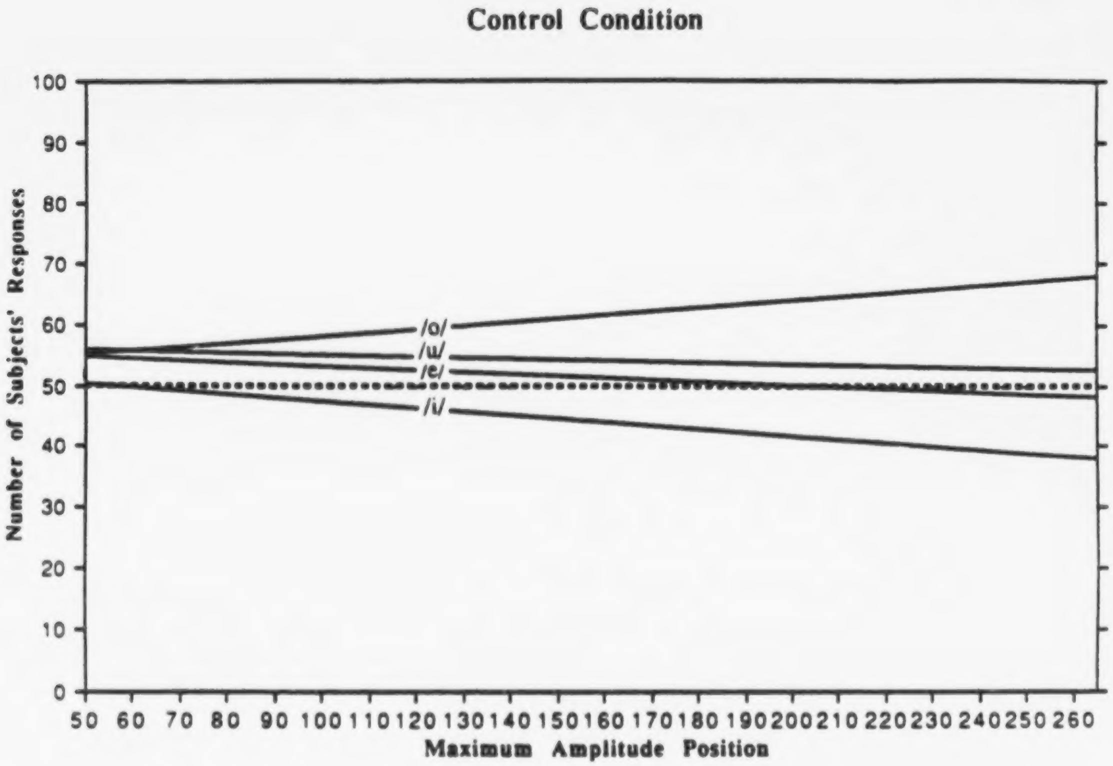


Figure 4. Lines of best fit for the four vowels in the control condition, plotting the percent of responses in which subjects perceived the first member of a vowel pair as more stressed than the second member. For the control condition, the percentile remaining above each line of best fit in this figure is the percent of responses in which subjects perceived the second member of a vowel pair as more stressed than the first member. The choice to plot responses for the first member rather than second member of vowel pairs for the control condition was arbitrary.

Table 2

F-values for maximum amplitude position and vowel type at each of the ten levels of amount different and the two levels of direction different.

Amount Different	Maximum Amplitude Position		Vowel Type	
Control	0.54		9.19***	
	Direction Different			
	A<B	A>B	A<B	A>B
10 msec	0.77	0.92	16.74 ***	9.32 ***
15 msec	1.24 +	1.64 **	68.42 ***	56.17 ***
20 msec	1.47 *	3.08 *	18.72 ***	25.98 ***
25 msec	0.98	1.28 ++	0.93	0.43
30 msec	1.00	0.82	19.43 ***	19.28 ***
35 msec	1.18 +	1.60 +	9.70 ***	13.32 ***
40 msec	1.24 +	0.96	13.27 ***	10.72 ***
45 msec	1.07	1.26 +	11.28 ***	5.46 ***
50 msec	1.27 +	2.68 ***	24.08 ***	6.26 ***

+	p>0.25
++	p>0.10
*	p>0.05
**	p>0.01
***	p>0.001

Insert Figures 5 through 12 about here

How great must the maximum amplitude position difference be in two successive syllables for a stress difference to be perceived?

As was mentioned above, based on the production task in Behne (1989), a 35 msec position difference is expected to perceptually distinguish stressed and unstressed syllable. If an early amplitude peak provides a perceptual cue to stress, an amount different of approximately 35 msec should be a large enough difference for perception.[6] The F-values for the levels of amount different show that stress perception is affected by maximum amplitude position differences as small as 15 msec for $A>B$ pairs [$F(40,1476)=1.64$; $p>0.01$], and 20 msec for $A<B$ pairs [$F(39,1440)=1.47$;$p>0.05$]. Maximum amplitude position differences between 25 msec and 45 msec do not clearly appear to be influencing stress perception; this point will be investigated further with more detailed analysis of the means across the maximum amplitude positions.

Does the maximum amplitude position affect stress perception of different vowels in the same manner?

As has been mentioned, stress perception of different vowels varied in the control group. The F-values in Table 2 reveal that stress perception also varied among the four vowels in the experimental levels of amount difference. A comparison of the vowels for the delimited data shown in Figures 5-12 further demonstrates a difference among the vowels. One outstanding characteristic in these figures is the tendency for /u/ to behave unlike the other vowels; although not clearly explainable at this point, the results seem to reflect subjects' comments that /u/ trials seemed particularly difficult. Although stress perception varying across vowel type does not appear to be systematically associated with vowel characteristics such as height or frontedness/rounding, a closer look at the stimuli and means will be necessary.

---

[6] Unlike the stimuli in the present study, the stressed and unstressed syllables in the production research of Behne (1989) were not sequential. For stressed and unstressed syllables occurring sequentially, a smaller amount different than 35 msec might be expected.

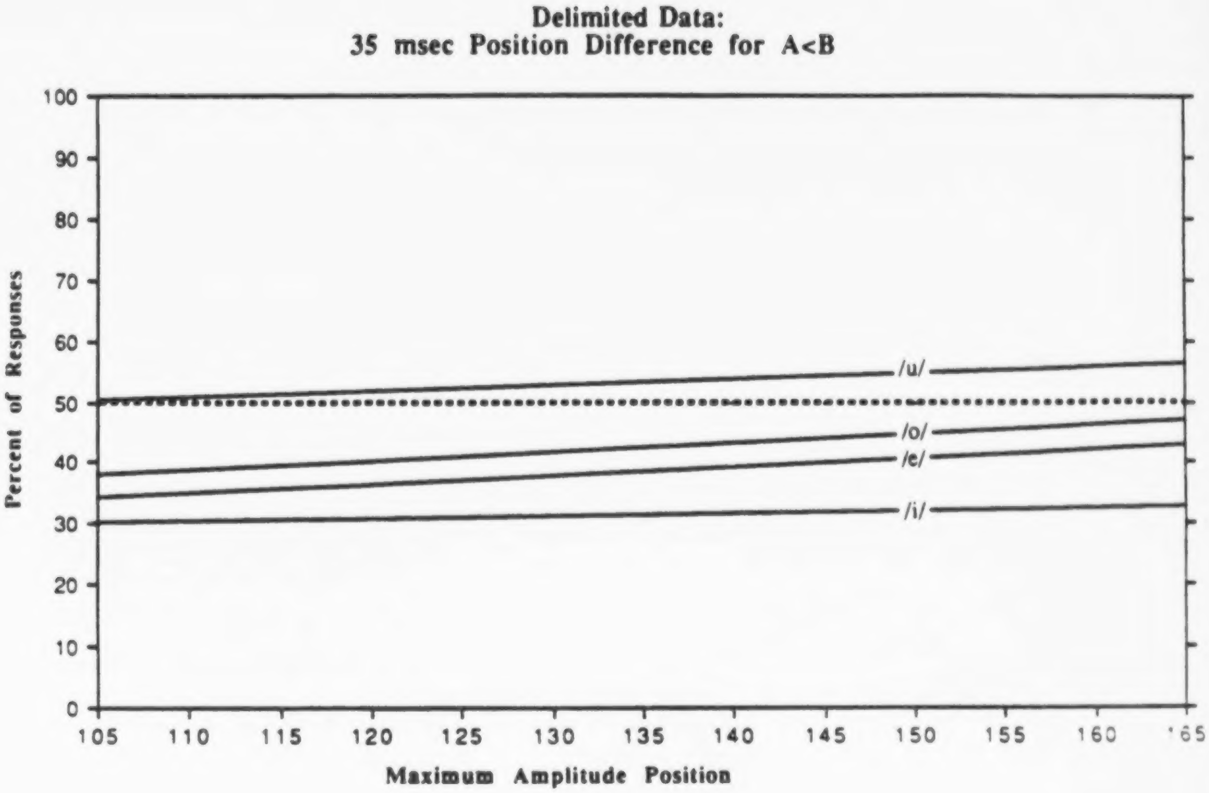


Figure 5. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the first member of a vowel pair as more stressed than the second member when the maximum amplitude position of the first member is 35 msec. earlier than that of the second member.

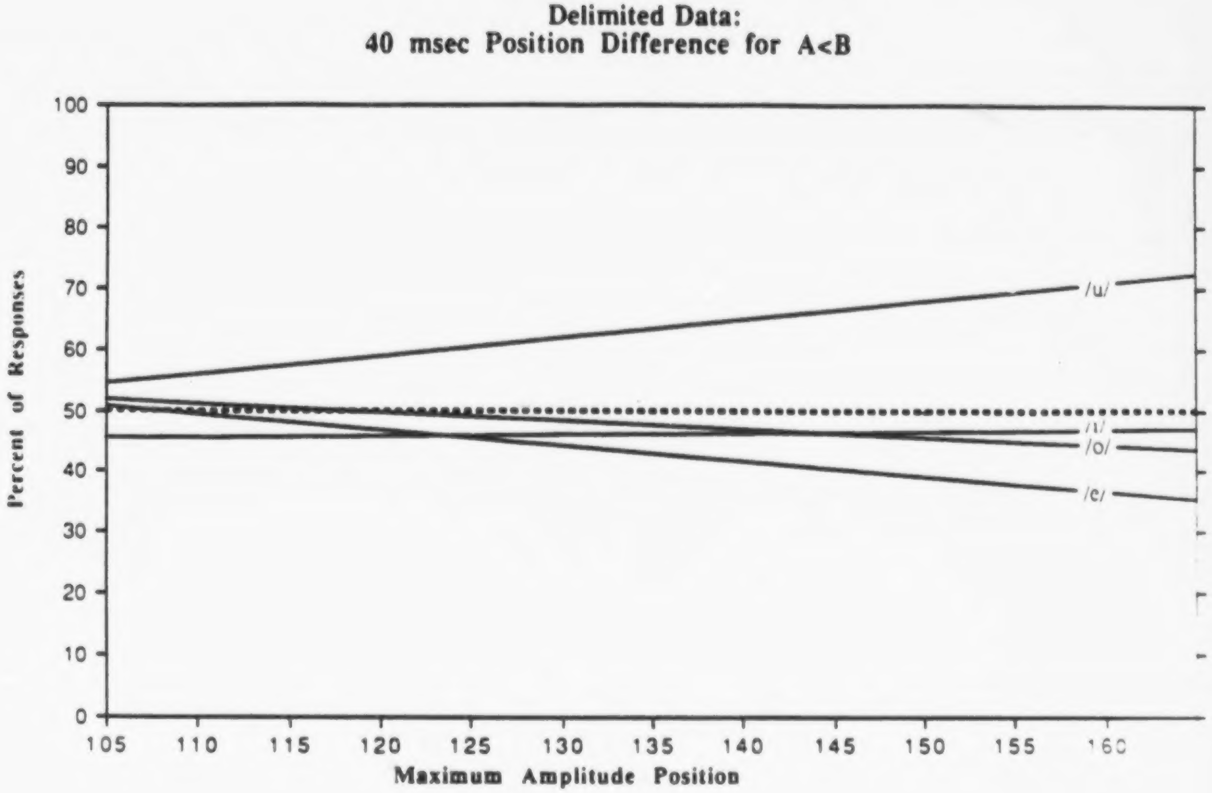


Figure 6. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the first member of a vowel pair as more stressed than the second member when the maximum amplitude position of the first member is 40 msec. earlier than that of the second member.

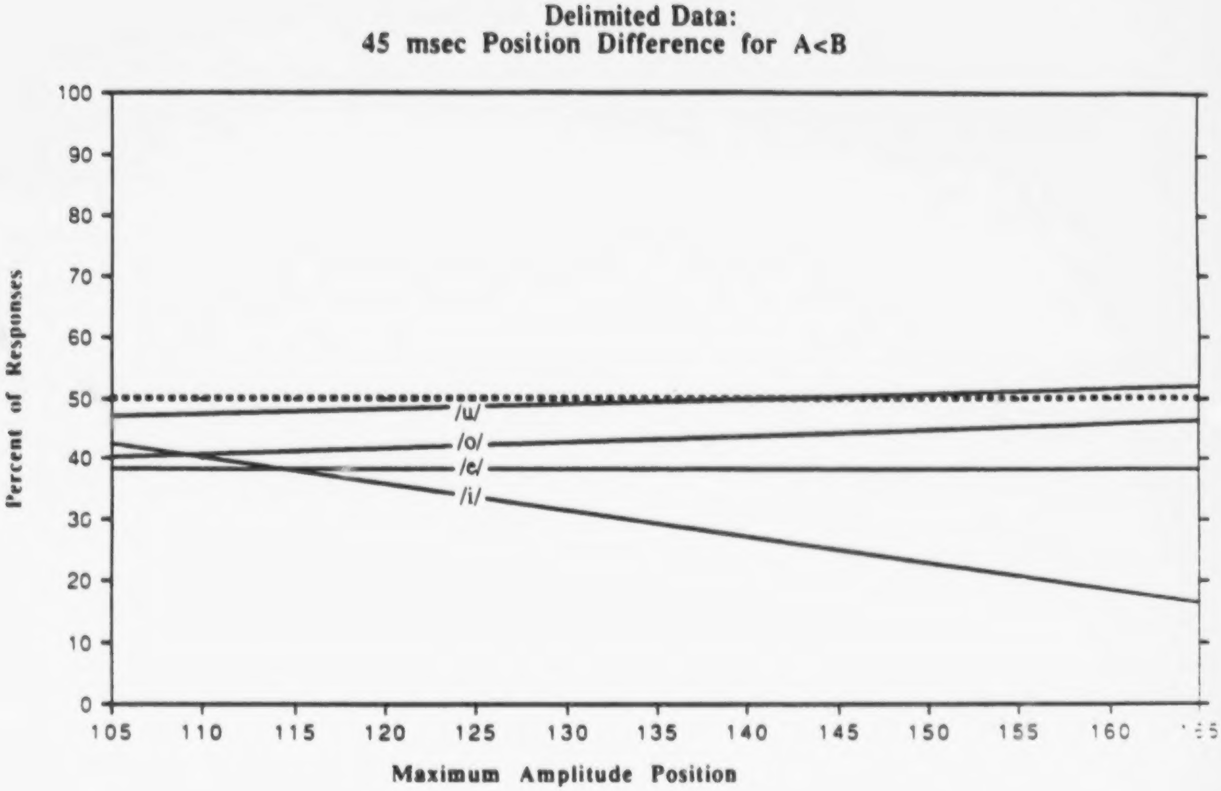


Figure 7. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the first member of a vowel pair as more stressed than the second member when the maximum amplitude position of the first member is 45 msec. earlier than that of the second member.

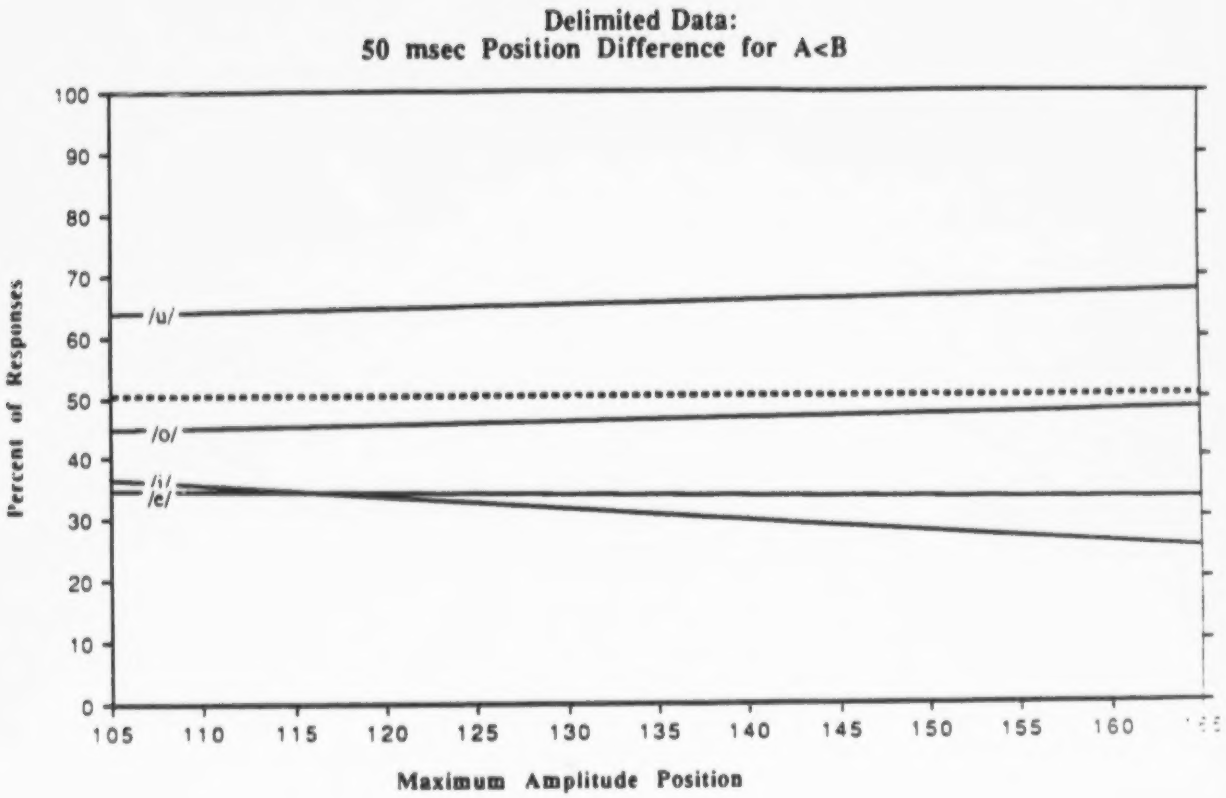


Figure 8. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the first member of a vowel pair as more stressed than the second member when the maximum amplitude position of the first member is 50 msec earlier than that of the second member.

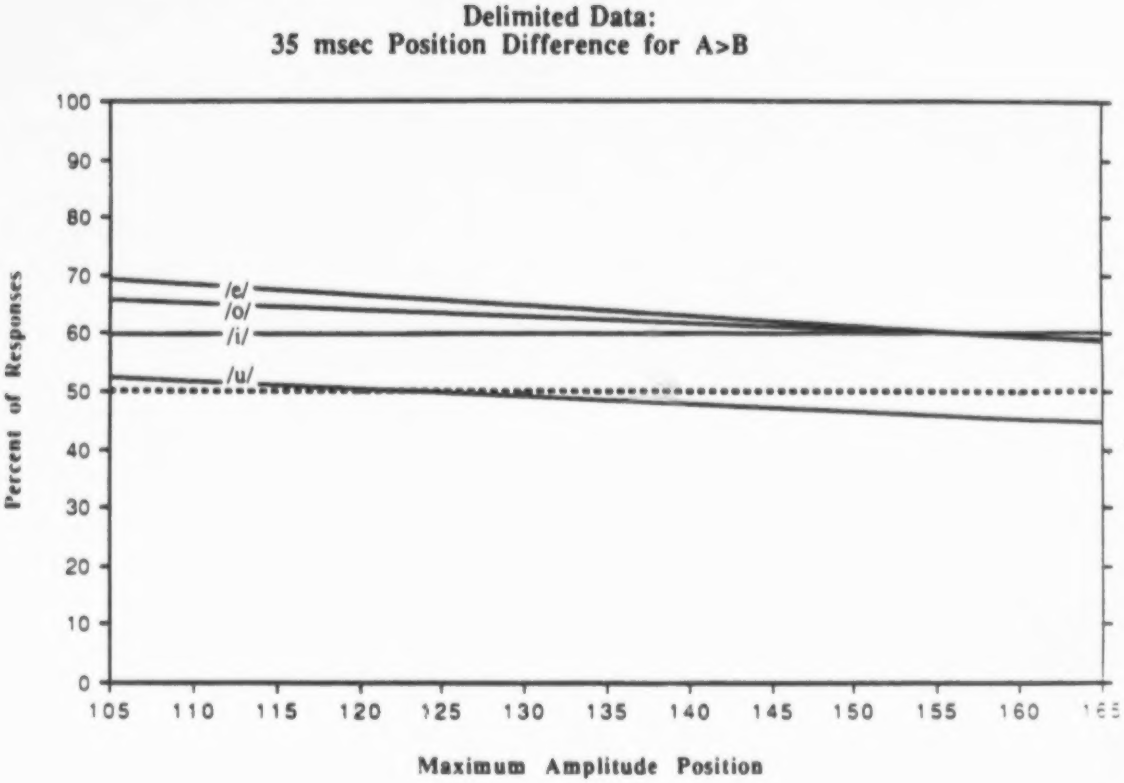


Figure 9. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the second member of a vowel pair as more stressed than the first member when the maximum amplitude position of the second member is 35 msec. earlier than that of the first member.

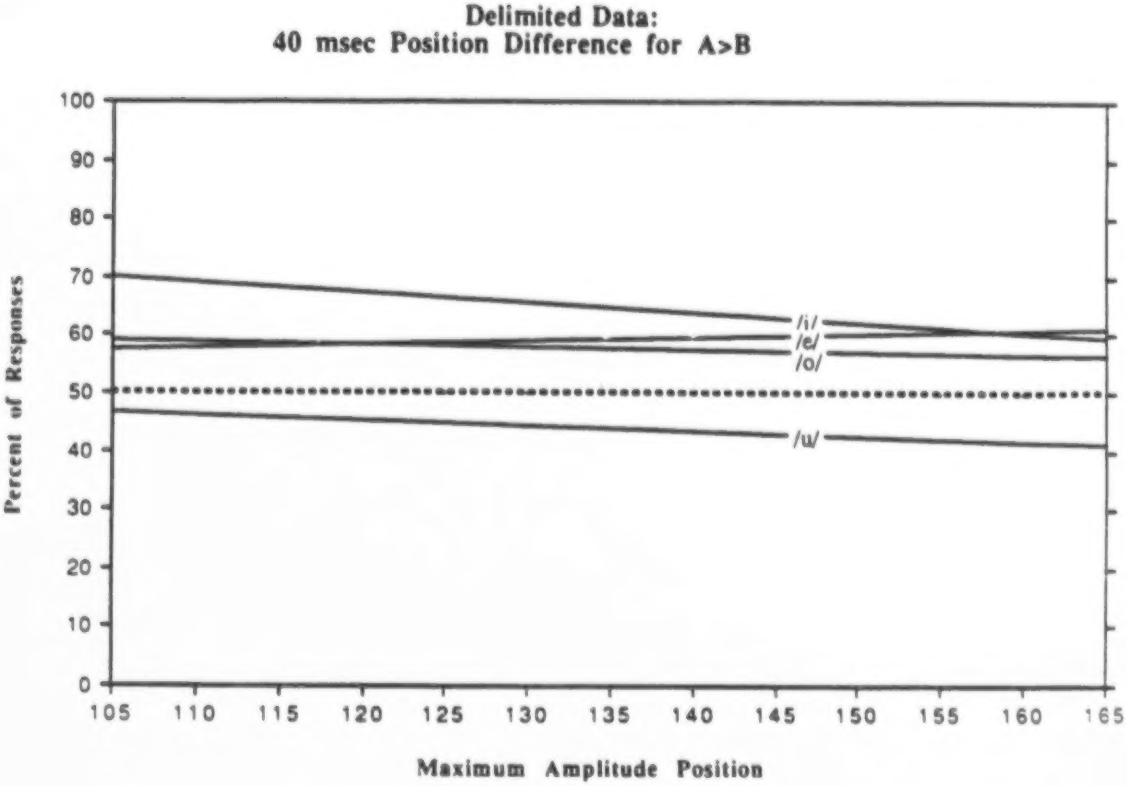


Figure 10. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the second member of a vowel pair as more stressed than the first member when the maximum amplitude position of the second member is 40 msec. earlier than that of the first member.

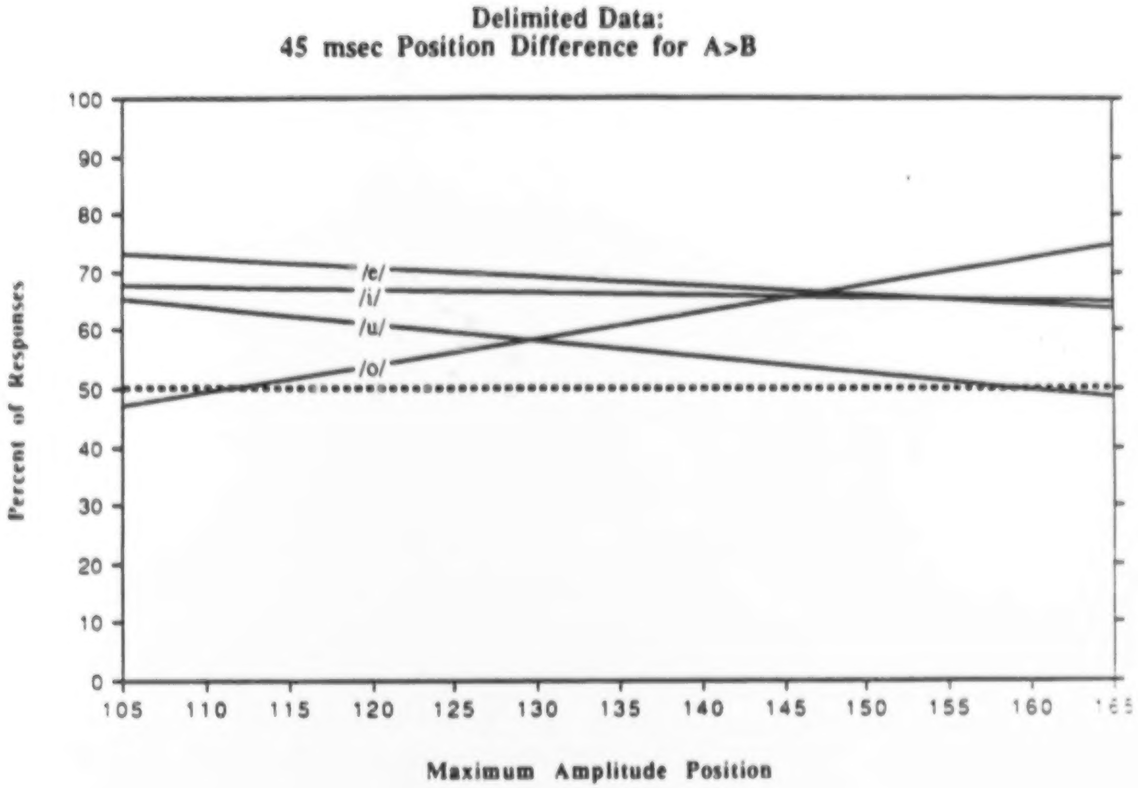


Figure 11. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the second member of a vowel pair as more stressed than the first member when the maximum amplitude position of the second member is 45 msec. earlier than that of the first member.

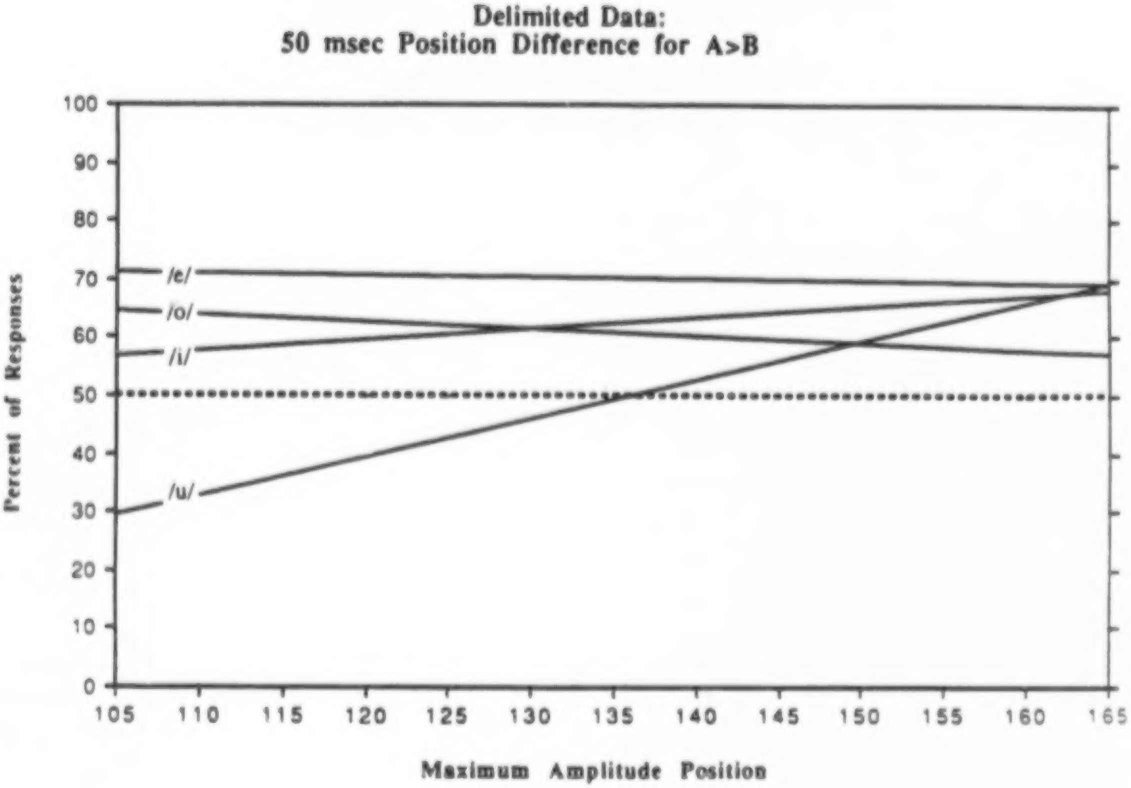


Figure 12. Lines of best fit for the delimited data of the four vowels, plotting the percent of responses in which subjects perceived the second member of a vowel pair as more stressed than the first member when the maximum amplitude position of the second member is 50 msec. earlier than that of the first member.

Is an early maximum amplitude able to serve as a perceptual cue for stress?

For the ten levels of amount different and two levels of direction different, the F-values of the maximum amplitude position factor are presented in Table 2. The results in Table 2 do not overwhelmingly show evidence that an early maximum amplitude provides a strong acoustic cue for stress; that is, although the F-values for the maximum amplitude position are not consistently significant for an amount different which is greater than at least 35 msec for either A<B or A>B vowel pairs, they suggest that stress perception is being influenced by the position of the maximum amplitude. In addition, the delimited data presented in Figures 5-12, especially for A>B vowel pairs, show a tendency toward response differences due to the early position of the maximum amplitude.

In summary, the present study was designed to provide an initial investigation of the maximum amplitude position as a stress cue in perception. To date, the data tend to support the hypothesis based on previous production research, that an early maximum amplitude is associated with stress. The maximum amplitude position appears to influence stress perception in such a way that a position difference of 15 msec may be large enough for a stress difference to be perceived. However, the the relative maximum amplitude positions in successive syllables appears to differentially affect stress perception. Furthermore, the perception of stress as a function of maximum amplitude position appears to be dependent on vowel type. Two points should be emphasized: First, the results have only been partially analysed at this point and conclusions will be held until the data have been more fully addressed. Second, this study is an initial investigation of the maximum amplitude position as a stress cue in perception and, consequently, generalizations of the results to natural speech are limited at this point. Nevertheless, the exploratory nature of this study and further analysis of the data hopefully will provide insight into the acoustic nature of stress.

References

Armstrong, L.E. (1932). *The phonetics of French.* London: G. Bell & Sons.

Behne, Dawn M. (1989). *Acoustic effects of focus and sentence position on stress in English and French.* Unpublished Doctoral Dissertation: University of Wisconsin-Madison.

Bolinger, Dwight. (1958). "A theory of pitch accent in English". *Word* **14**, 109-149.

Bloomfield, Leonard. (1933). *Language.* New York: Holt.

Coustenoble, H.N. & L.E. Armstrong. (1934). *Studies in French intonation.* Cambridge: W. Heffner & Sons.

Delattre, Pierre. (1963). "Comparing the prosodic features of English, German, Spanish and French." *International Review of Applied Linguistics* **4**, 183-198.

Fry, D.B. (1955). "Duration and intensity as physical correlates of linguistic stress". *The Journal of the Acoustical Society of America* **27**, 765-768.

Fry, D.B. (1958). "Experiments in the perception of stress". *Language & Speech* **1**, 126-152.

House, Arthur & Grant Fairbanks. (1953). "The influence of consonant environment upon the secondary acoustical characteristics of vowels" *The Journal of the Acoustical Society of America* **25**, 105-113.

Jones, D. (1960). *An outline of English phonetics*, (9th ed). Cambridge: W. Heffer & Sons.

Ladefoged, Peter. (1967). *Three areas of experimental phonetics.* Oxford: Oxford University.

Lehiste, Ilse. (1970). *Suprasegmentals.* Cambridge, MA: MIT.

Lieberman, Philip. (1960). "Some acoustic correlates of word stress in American English", *The Journal of the Acoustic Society of America* **32**, 451-454.

Morton, John & Wiktor Jassem. "Acoustic correlates of stress", *Language & Speech* **8**, 159-181.

Oller, D.Kimbrough. (1973). "The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America* **54**, 1235-1247.

Paramenter, C.E. & S.N. Treviño. (1935). "The length of the sounds of a Middle Westerner", *American Speech* **10**, 129-133.

Passy, Paul. (1907). *The sounds of the French language.* Oxford: Clarendon.

Stetson, R.H. (1928). *Motor phonetics.* Amsterdam: North Holland.

Sweet, H. (1890). *A Primer in phonetics.* Oxford: Clarendon.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

A Comparison of the First and Second Formants of Vowels Common to English and French[1]

Dawn M. Behne

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported, in part, by NIH Training Grant No. NS-07134-11 to Indiana University in Bloomington, IN. The author would like to thank W. Charles Read of the University of Wisconsin-Madison for comments on an earlier version of the paper.

Abstract

Delattre (1965) has characterized English vowels as being lower, more central, and less rounded than the predominantly high, front, rounded French vowels. In the present study, Delattre's comparison is experimentally investigated by comparing the first and second formant frequencies of English and French vowels which occur in both languages. The results suggest that (1) high vowels tend to be higher in French than in English, (2) /ɑ/tends to to be lower in English than in French, (3) high and mid vowels tend to be more central in English than French, (4) rounded vowels tend to have greater lip rounding in French than in English, (5) front vowels are higher and more fronted in French than in English, and (6) back vowels have greater pharyngeal constriction in English than in French. Although Delattre's description of English and French vowels is generally supported by the results, it does not fully the characterize the differences demonstrated between the vowels common to English and French.

A Comparison of the First and Second Formants of Vowels Common to English and French

Delattre (1965) has described the English vowels as being lower, more central, and less rounded than the predominantly high, front, rounded vowels of French. Specifically, he describes the following differences:

1. The high vowels are lower in English than in French.
2. The low vowels are lower in English than in French.
3. The high and mid vowels are more central in English than in French.
4. The rounded vowels are less rounded in English than in French.

Although Delattre's comparison of English, French, German and Spanish was an experimental investigation, the procedure used in reaching the characterization of vowel quality in English and French is not explicitly stated. In the present study, Delattre's description is reconsidered experimentally by comparing the F1 and F2 of vowels common to English and French, /i, e, ɛ, u, o, ɔ, ɑ/.

To investigate the relative height, centrality and roundedness of English and French vowels, Delattre's comparison can be restated in terms of F1 and F2 relationships which are summarized in Table 1.

Insert Table 1 about here

First, Delattre's comparison addresses two points of vowel height: (a) the high vowels, /i/ and /u/, are higher in French than in English, and (b) the low vowel, /ɑ/, is higher in French than in English. Since F1 is lowered by front oral cavity constriction and raised by pharyngeal constriction, the F1 of /i, u, ɑ/ ought to be lower in French than in English. The frequency of F2 is raised by front tongue constriction and lowered by back tongue constriction; consequently, the front vowel, /i/, ought to have a higher F2 in French than in English, and the back vowel, /u/, ought to have a lower F2 in French than in English.

Second, Delattre describes French vowels as being more extreme than English vowels; specifically, the high and mid vowels, /i, e, ɛ, u, o, ɔ/, are characterized as being more central in English than in French. Because F2 becomes higher with increased front tongue constriction, the F2 of the high and mid front vowels, /i, e, ɛ/, ought to be higher in French than in English. The increased back tongue constriction which lowers F2, ought to result in the F2 of the high and mid back vowels, /u, o, ɔ/, being lower in French than in English.

Finally, Delattre describes vowels as being more rounded in French than in English. Since the formant frequencies of both F1 and F2 tend to be lower with increased lip rounding, F1 and F2 of the rounded vowels, /u, o, ɔ/, are expected to be lower in French than English.

Table 1

The expected relative formant frequencies of English and French based on Delattre's description.("—" signifies that the relationship was not predictable from Delattre's description.)

	F1	F2
/i/	French < English	French > English
/e/	—	French > English
/ɛ/	—	French > English
/u/	French < English	French < English
/o/	French < English	French < English
/ɔ/	French < English	French < English
/ɑ/	French < English	—

These F1 and F2 expectations are investigated in order to compare the relative height, centrality and roundedness of vowels common to English and French.

Method

Subjects. Fourteen monolingual speakers of American English and fourteen monolingual speakers of French participated in the project.[1] In order to limit inter-speaker variance as much as possible, all participants were men between twenty and fourty-five years old.

Stimuli. Stimuli were developed using the seven distinctive vowels common to English and French: /i, e, ɛ, u, o, ɔ, ɑ/. Each vowel was contained in a real, monosyllabic English word and French word. All vowels except /e/ were nested between two consonants. Since /e/ occurs only in open syllables in French monosyllabic words, /e/ was used in an open syllable in both the French and the English words in order to maintain parallelism between the stimuli for the two languages. The consonants surrounding the target vowels in both languages were limited to the unvoiced alveolar obstruents /t/ and /s/ which (a) allowed real words to be used for all target vowels in both languages, and (b) controlled for variation in formant frequencies resulting from differences in voicing and place of articulation. When possible, the consonant environment for a particular vowel was identical for the two languages. The English and French stimuli for the seven vowels are listed in Table 2.

Insert Table 2 about here

Procedure. A recording was made of each American subject saying each English word in the context "Say ___", and of each French subject saying each French word in the context "C'est ___", meaning "This is ___". These environments were chosen because they provide a semantically neutral carrier sentence with a phonetically similar (i.e. [se]) context.

The F1 and F2 of each subject's production of the seven vowels were measured using the autocorrelation method of LPC (14 coefficients) and peak picking method of formant estimation (both from ILS software package). To ensure that the formant frequencies being collected were from the steady state of the vowel, the formant frequencies were measured near the center of each vowel.

---

[1] All of the American subjects were originally from the midwest and French subjects were from a variety of regions in France.

Table 2

English and French stimulus words in which the seven target vowels were tested.

	ENGLISH		FRENCH	
/i/	seat	/sit/	cite	/sit/
/e/	say	/se/	the	/te/
/ɛ/	set	/sɛt/	cette	/sɛt/
/u/	toot	/tut/	toute	/tut/
/o/	tote	/tot/	saute	/sot/
/ɔ/	saught	/sɔt/	sotte	/sɔt/
/ɑ/	tot	/tɑt/	tate	/tɑt/

Results and Conclusions

In order to avoid distorting interlanguage differences and similarities with inter-speaker variance of a language, analysis of variance was used to compare the vowel patterns of English and French .[2] Analysis of variance allows variance among vowels common to English and French and spoken by different subjects to be broken down so that the source, or sources, of the variance can be systematically identified. Separate two-factor analyses of variance were conducted for the F1 and F2 of the target vowels in English and French. The mean formant frequencies for each vowel and for each language are presented in Table 3 and displayed in Figure 1. The mean formant frequencies for front, back and rounded vowels are presented in Table 4.

Insert Tables 3, 4 & 5, and Figure 1 about here

First Formant. The average frequency of F1 in English vowels ($\bar{x}$ =512Hz) is greater than in French vowels ($\bar{x}$ =419Hz), [$F(1,26)=73.91$; $p<0.001$].[3] Separate analyses of F1 for front and back vowels demonstrated a higher F1 for front vowels (i.e. less oral constriction) in English ($\bar{x}$ =451Hz) than in French ($\bar{x}$ =373Hz), [$F(1,26)=61.86$; $p<0.001$], and a higher F1 for back vowels (i.e. greater pharyngeal constriction) in English ($\bar{x}$ =559Hz) than in French, ($\bar{x}$ =454),[$F(1,26)=59.91$; $p<0.001$]. Rounded vowels, all of which are included in the group of back vowels, were also found to have a significantly higher F1 in English ($\bar{x}$ =493) than in French ($\bar{x}$ =390), [$F(1,26)=59.53$; $p<0.001$].

Second Formant. The average frequency of F2 was not significantly different for for English and French vowels [$F(1,26)=3.9$; $p<0.10$]. Separate analyses of front, back and rounded vowels compared English and French F2 frequencies. Front vowels had a higher F2 (i.e. greater front tongue constriction) in French ($\bar{x}$ =1912Hz) than in English ($\bar{x}$ =1813), [$F(1,26)=4.9$; $p<0.05$]. No significant F2 differences were found between English and French back[$F(1,26)=0.26$; n.s.] or rounded [$F(1,26)=0.00$; n.s] vowels.

Vowel Height, Centrality and Roundedness. In order to identify specific differences addressed by Delettre's comparison of English and French vowels, analyses of variance for F1 and for F2 compared each vowel across English and French. The results of the analyses of

---

[2] Disner (1986) points out that the normalization procedures intended to extract between speaker variance in many studies of vowel quality use a correction factor which is not appropriate for comparing vowel systems across languages.

[3] The significance level of 0.05 is accepted as standard for this investigation. Although not accepted as significant, F-values at $0.10<p<0.05$ are noted, and other nonsignificant F-values are marked "n.s."

Table 3

Mean F1 and F2 frequencies for the seven target English and French vowels.

	ENGLISH		FRENCH	
	F1	F2	F1	F2
/i/	310	2095	262	2043
/e/	464	1799	364	1969
/ɛ/	579	1546	494	1724
/u/	340	1237	284	1062
/o/	504	944	397	939
/ɔ/	635	1105	488	1277
/ɑ/	755	1332	646	1409
$\bar{x}$ =	512	1437	419	1489

Table 4

Mean F1 and F2 frequencies for front /i, e, ɛ/, back /u, o, ɔ, ɑ/ and rounded /u, o, ɔ/ English and French vowels.

	ENGLISH		FRENCH	
	F1	F2	F1	F2
Front	451	1813	373	1912
Back	559	1155	454	1172
Rounded	493	1095	390	1093

Table 5

The relative formant frequencies of English and French. (Capitals indicate that the results concur with Delattre's (1965) description; "—" signifies that the relationship was not predictable from Delattre's description.)

	F1	F2
/i/	FRENCH < ENGLISH	n.s
/e/	—	FRENCH > ENGLISH
/ɛ/	—	FRENCH > ENGLISH
/u/	FRENCH < ENGLISH	FRENCH < ENGLISH
/o/	FRENCH < ENGLISH	n.s.
/ɔ/	FRENCH < ENGLISH	French > English
/ɑ/	FRENCH < ENGLISH	—

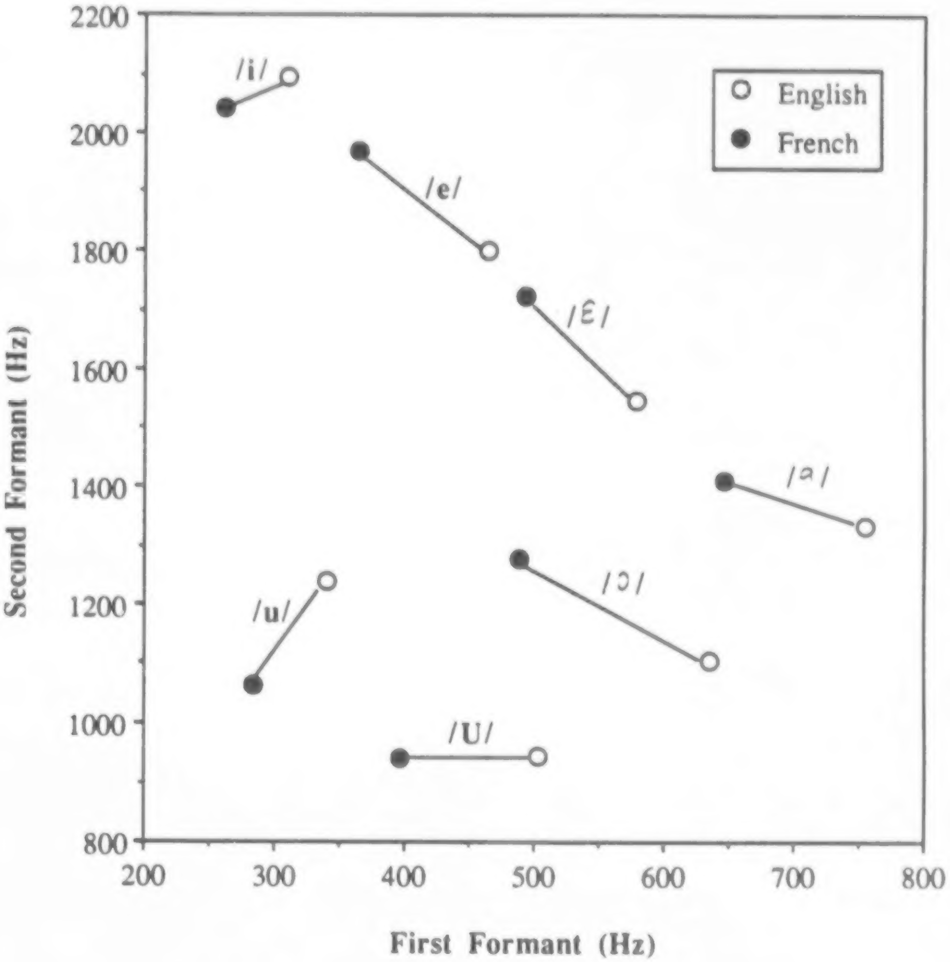


Figure 1. Display of the average F1 and F2 of English and French vowels.

variance for the means in Table 3 are presented in Table 5.

1. Are the high vowels lower in English than in French?

Greater tongue height lowers F1, raises the F2 of front vowels and lowers the F2 of back vowels. The English and French F1 frequencies for /i/ and /u/ were compared. For both /i/ [$F(1,26)=24.14$; $p<0.001$] and /u/ [$F(1,26)=14.31$; $p<0.001$], F1 was lower in French (/i/,$\bar{x}$ =262;/u/,$\bar{x}$ =284) than in English (/i/,$\bar{x}$ =310;/u/,$\bar{x}$ =340). Comparing the English and French F2 frequencies for /i/ and /u/, no significant difference was found between the F2 of /i/ in English ($\bar{x}$ =2095Hz) and French ($\bar{x}$ =2043Hz), [$F(1,26)=0.57$; n.s.], but the F2 of /u/ was significantly lower in French ($\bar{x}$ =1062Hz) than in English ($\bar{x}$ =1237Hz), [F91,26)=4.87; $p<0.05$]. The results indicate that /u/ is, and /i/ tends to be, qualitatively higher in French than in English.

2. Is /ɑ/ lower in English than in French?

Lowering a back vowel, such as /ɑ/, results in a higher F1 associated with pharyngeal constriction and a lower F2 due to back tongue constriction. The mean F1 for /ɑ/ is significantly higher [$F(1,26)=19.49$; $p<0.001$] in English ($\bar{x}$ =755Hz) than in French ($\bar{x}$ =646Hz). Although the F2 of /ɑ/ is lower in English ($\bar{x}$ =1332Hz) than in French ($\bar{x}$ =1409Hz), the difference was not statistically significant [$F(1,26)=4.22$; $p<0.10$]. As a whole, the results suggest that /ɑ/ is qualitatively lower in English than in French.

3. Are the high and mid vowels more central in English than in French?

The frequency of F2 rises when front tongue constriction increases, and lowers when back tongue constriction increases. A comparison of the English and French F2 frequencies for each of the high and mid vowels shows:

(a) no difference between the F2 of/i/ in English and in French [$F(1,26)=0.57$; n.s];

(b) the F2 of /e/ is higher in French ($\bar{x}$ =1969Hz) than in English ($\bar{x}$ =1799) [$F(1,26) = 9.36$; $p<0.01$];

(c) the F2 of /ɛ/ is higher in French ($\bar{x}$ =1724Hz) than in English ($\bar{x}$ =1546) [$F(1,26) = 33.72$; $p<0.001$];

(d) the F2 of /u/ is lower in French (x =1062) than in English ($\bar{x}$ =1237Hz) [$F(1,26) = 4.87$; $p<0.05$];

(e) although, the F2 of /o/ is slightly lower in French ($\bar{x}$ =939) than in English ($\bar{x}$ =944), the difference was not significant [$F(1,26)=0.01$; n.s.];

(f) the F2 of /ɔ/ is higher in French ($\bar{x}$ =1277Hz) than in English ($\bar{x}$ =1105) [$F(1,26) = 23.51$; $p<0.001$].

The vowels /e, ɛ, u, ɔ/ appear to be more central in English than in French, but /i/ and /o/ do not clearly follow the pattern of the other high and mid vowels.

4. Are the back round vowels less rounded in English than in French?

Lip rounding is associated with a general lowering of formant frequencies. A comparison of English and French showed that the F1 of rounded vowels is significantly lower [$F(1,26)=59.53$; $p<0.001$] in French($\bar{x}$ =390Hz) than in English ($\bar{x}$ =493Hz). However, no significant difference was found between the mean English and French F2 frequencies [$F(1,26)=0.00$; n.s.]. Investigation of individual rounded vowels demonstrated a significantly lower F1 for /u/ [$F(1,26)=14.31$; $p<0.001$], /o/ [$F(1,26)=25.16$; $p<0.001$] and /ɔ/ [$F(1,26)=26.45$; $p<.001$] in French (/u/, $\bar{x}$ = 284Hz; /o/, $\bar{x}$ =397Hz; /ɔ/, $\bar{x}$ =488Hz) than in English (/u/, $\bar{x}$ = 340Hz; /o/, $\bar{x}$ =504Hz; /ɔ/, $\bar{x}$ =635Hz). The F2 of /u/ was lower in French ($\bar{x}$ =1062Hz) than in English ($\bar{x}$ =1237Hz), [$F(1,26)=4.87$; $p<0.05$], but higher for /ɔ/ in French ($\bar{x}$ =1277Hz) than in English ($\bar{x}$ =1105Hz), [$F(1,26)=23.51$; $p<0.001$]. No significant difference was shown between English and French for the F2 of /o/ [$F(1,26)=0.01$; n.s.]. The combined results of F1 and F2 suggest a tendency for the back, rounded vowels to be more rounded in French than in English.

In summary F1 and F2 measurements of the vowels common to English and French were used to reconsider Delattre's comparison of English and French vowels. The analyses of height, centrality and roundedness for /i, e, ɛ, u, o, ɔ, ɑ/ lead to the following conclusions:

1. The high vowels are generally higher in French than in English.
2. The low back vowel tends to be lower in English than in French.
3. High and mid vowels tend to be more central in English than in French.
4. Back rounded vowels tend to have less lip rounding in English than in French.

Beyond addressing Delattre's descriptions of English and French vowels, the results indicate that front vowels are produced with more front oral constriction and front tongue constriction in French than in English, and that back vowels are produced with greater pharyngeal constriction in English than in French. Although these conclusions generally support Delattre's comparison of English and French vowels, the systematic relationships Delattre suggested do not fully characterize the differences demonstrated between the English and French vowels. A comparison of the vowels common to English and French appears to be somewhat more complex than Delattre proposed.

References

Delattre, Pierre. (1965). *Comparing the phonetic features of English, French, German and Spanish: An interim report.* Heidelberg: J. Groos Verlag.

Disner, Sandra. (1986). "On describing vowel quality". In J. Ohala & J. Jaeger (eds), *Experimental phonology.* Orlando: Academic.

Age Differences in Spoken Word Identification: Effects of Lexical Density and Semantic Context[1]

Theodore S. Bell[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The author acknowledges valuable assistance from David Pisoni, Larry Humes, Keith Johnson, Judith Hennessey, Heidi Banholzer, Samantha Way, Ingrid Nobel and Michael Turley.

[2] Theodore Bell is on leave from the UCLA School of Medicine, Department of Surgery, Division of Head and Neck, Los Angeles, CA 90024

Abstract

Young and elderly listeners were compared on word identification performance in noise as a function of target word frequency, phonetic similarity neighborhood, and degree of semantic context provided by the carrier sentence. The 10 elderly listeners had pure-tone thresholds better than 65 dB HL at frequencies below 4000 Hz and all were older than 65 years of age. The young listeners were 10 normal-hearing undergraduate volunteers between 18 and 32 years of age. Sentences were presented monaurally (85 dB SPL) over earphones in speech-peak shaped noise at -2 dB S/N; the task was to identify the last word of each spoken sentence. Each target word was categorized on the basis of word frequency and its phonetic similarity to other words in the lexicon, and each word was presented in two sentence frames differing in terms of semantic context. While semantic context was always beneficial, the elderly group showed a larger word frequency effect in the presence of semantic context and a reduced effect in the absence of context relative to the younger listeners. Some age-related changes associated with receptive speech communication may reflect an increased reliance on semantic and lexical information as compensation for degraded peripheral and central encoding. A word frequency by phonetic similarity interaction also was evident in the data, indicating agreement with the Neighborhood Activation Model, and further demonstrating that word frequency effects occur after acoustic-phonetic pattern recognition.

Age Differences in Spoken Word Identification: Effects of Lexical Density and Semantic Context.

Studies of speech reception difficulties of elderly and hearing-impaired listeners may give insight into some longstanding issues central to the topic of speech perception in general. For example, a critical issue within the speech perception literature has been the specification of peripheral and central encodings and their influence on word recognition and identification processes. In the gerontology literature, this same problem, the separation of sensory and cognitive influences, has become a central issue, a prerequisite for understanding aging and its accompanying deficits in speech perception and language comprehension.

Difficulty with receptive speech communication is a major complaint among the elders of our population. Studies of communication deficits of older persons have increased in recent years, partly due to demographic forecasts indicating that the growing number of elderly in our society will strain social and health-related resources as we enter the next century. The Bureau of the Census predicts that by the year 2030 the number people over 65 years of age will be 52 million, or up to 21 percent of the total population. Hearing deficits associated with aging are often confounded with other age-related deficits, and as a result, the loss of speech receptive abilities may not be attributed exclusively to sensory deficits; often there are underlying central etiologies (e.g., Bergmann, Blumefield & Levitt, 1976; Ford & Roth, 1977; Hayes & Jerger, 1979; Bosatra & Russalo, 1982; Weinstein & Ventry, 1983; Welsh, Welsh & Healy, 1985; Jerger, et al., 1989). Thus, to tackle this problem effectively it is necessary to gain a fundamental understanding of how speech is processed by young and old alike, including the nature of acoustic-phonetic encodings, lexical decision processes, and the role of syntactic and semantic context comprehension.

The elderly listener has several important auditory and cognitive characteristics that are distinct from the young normal-hearing listener typically employed in basic speech research. The peripheral auditory system of an elderly listener is likely to be impaired from a lifetime of acoustic insults; a partial list would certainly include noise exposure in the home and workplace, as well as exposure to ototoxic drugs (cf. Lowell & Paparalla, 1977; Butler & Gastel, 1979; Cervellera & Quaranta, 1982). Elderly adults also have a knowledge base built up over a lifetime that allows them to compensate for the slowing of cognitive processes in general (Lorsbach & Simpson, 1983; Wingfield, Poon, Lombardi & Lowe, 1985; Roedder & Cole, 1986) as well as the lack of stimulus specificity caused by sensory degradation (e.g., West & Cohen, 1985; Stinson & Tracy, 1983; Spilich & Voss, 1983; Cohen & Faulkner, 1983). Nonacoustic knowledge sources have been known for some time to also contribute to the identification of words in normal continuous discourse (e.g., Marlsen-Wilson & Tylor, 1980; Salasoo & Pisoni, 1985). Thus, studies of the elderly listener may offer unique insights into the speech perception mechanism itself.

Before continuing, several terms that will be used throughout the following report require definition. Using the conventions of Salasoo & Pisoni (1985), the term "word identification" refers to the listener's understanding, or belief, whereas "word recognition" refers to lower order acoustic-phonetic pattern matching processes. "Lexical access" refers to the retrieval or activation of an item in working memory by connecting an internal representation derived directly from the speech input with its associated lexical representation in memory.

It may not be surprising that semantic context provides additional information available to the listener for use in deciphering the speech code; however, exactly how and when this information is employed is not completely understood. Marlsen-Wilson & Tyler (1980) found that listeners needed less than half of the acoustic-phonetic code in order to understand words in normal sentence contexts. Further support comes from studies that have used a stimulus gating paradigm to measure the minimum acoustic-phonetic input required for word identification (Grosjean, 1980; Cotton & Grosjean, 1984; Salasoo & Pisoni, 1985). These studies showed that less stimulus information was required to identify words in sentences compared to identification of the same words in isolation. Further, the initial word segments were more important than ending segments, and were interpreted as reflecting a shift in information processing from bottom-up to top-down within the word identification process itself.

Grosjean (1980, 1985) found that incorrect responses in the gating task included not only acoustically similar words, but semantically related words as well. These data were used by Grosjean as evidence against the claim that only acoustic-phonetic information was used to compose the set of possible lexical candidates. His conclusion was that a model similar to Morton's (1979) interactive logogen model was required to explain these data, where both acoustic and nonacoustic knowledge sources interact to select a possible word candidate.

A recent study by Salasoo & Pisoni (1985) has provided support for Marlsen-Wilson's "principle of bottom-up priority." Acoustic-phonetic patterns are the primary source of information used to form a set of lexical candidates accessible from long-term memory, although semantic and syntactic information available from sentence contexts also provide additional candidates to the pool of potential words. The balance between these sources of knowledge in bottom-up and top-down processes allow the listener to comprehend speech even when the encoding is impoverished either by noise or sensory impairment.

The results of several recent studies from the Speech Research Laboratory at Indiana University have reduced the set of viable models in the literature to a subset of only a few, including the Neighborhood Activation Model (NAM) of Luce (1986). The NAM model captures many aspects of the word recognition process, and will serve as the basis of the present discussion.

Goldinger, Luce & Pisoni (1989) have recently provided additional data to support the Neighborhood Activation Model. The NAM assumes that the recognition of spoken words is characterized by a process in which phonetically similar words in memory initially are activated. Then, the member of the activated set that is most consistent with the acoustic-phonetic information in the speech waveform is selected. Further, it is assumed that word frequency biases responses toward the more likely, or frequent, members of the activated neighborhood (Luce, 1986; Luce, Pisoni & Goldinger, in press).

Assuming that the acoustic-phonetic code is degraded for the elderly listeners as compared to the younger listeners, the degraded stimulus leads to an inherently larger neighborhood of phonetically similar words. The impoverished sensory encoding of the elderly auditory systems leads to acoustic-phonetic encodings that are "fuzzy", resulting in greater similarity to more possible elements in the lexicon relative to a well-specified stimulus. Thus, one consequence of an impaired auditory system is that lexical similarity neighborhoods may be larger than they would otherwise be in a younger adult.

Word frequency effects should also be diminished for stimuli that are ill-defined. In the Neighborhood Activation Model, the size of the neighborhood would be dense, and although high-frequency words would be more likely, the ratio of stimulus frequency to neighborhood frequency would tend to be dominated by the neighborhood frequencies.

To summarize the predictions thus far, for a degraded stimulus encoding, the denominator term in the Neighborhood Activation Model becomes large, since it increases with increasing neighborhood size, leading to a decrease in the probability of correct identification of a stimulus word. In addition, frequency biases should become less evident as a result of poorly specified stimuli. If the number of lexical neighbors is large, then the relative influence of their combined frequency weightings dominates the ratio of stimulus to neighborhood frequency. Thus, the model predicts that a degraded acoustic-phonetic representation will not only lead to longer recognition times and less accurate identification, but to diminished frequency effects as well.

The purpose of the present study was to examine the perception of words embedded in sentences of varying degrees of semantic content. Both young and elderly listeners were used to examine interactions between lexical and semantic processes underlying speech perception. Previous studies have used young normal-hearing subjects and typically have employed words in isolation in quiet listening environments. The present study differs from earlier studies in several significant aspects. The stimuli were limited to CNC constructions embedded in sentence frames that provide varying degrees of semantic context, or predictability. Furthermore, the stimuli were presented in a spectrally shaped noise at a signal-to-noise ratio known to elicit performance levels in the range of 20 to 80 percent, dependent on the semantic content of the carrier sentence. Properties of the noise shaping are discussed below, but the arrangement is such that there should be no interaction with hearing sensitivity characteristics.

Method

Subjects. Subjects were 10 young adult listeners with normal hearing and 10 elderly hearing-impaired listeners with mild-to-moderate losses of unspecified etiology.

The young subjects were recruited through an introductory psychology class at Indiana University and received instructional credit. All of the young subjects were native speakers of English, and none reported any deficit in hearing sensitivity or had any history of auditory disfunction. The average age of the young listener population was 20.2 years.

The elderly listeners were recruited from a subject pool in the Speech and Hearing Sciences Department at Indiana University. All elderly subjects were paid for their participation in the study. The ages of the elderly listeners ranged from 65 to 83 years. Only subjects exhibiting less than a 70 dB (HL) loss of auditory sensitivity (re: ANSI S3.6-1969) at all frequencies below 4000 Hz were included in the study.

Stimuli. Stimulus target words were selected from the Lehiste & Peterson (1959) lists, which were arranged in 10 phonemically balanced lists, each with 50 words; only the first two lists were used in the present study. The distribution of phonemes in each list approximates the distribution of phonemes found in the English language generally, and frequency of occurrence of words in each list was representative of words in the language in general, and was comparable across lists as well.

Each target word was embedded as the final word in two distinct sentences. One sentence was constructed such that it gave no semantic indication of what the final word could be. The second sentence was composed in such a manner as to make the final word predictable. The same criteria as those employed by Kalikow et al. (1977) to generate the Speech Perception in Noise Test (SPIN) were used. Each sentence was approximately seven to nine words in length. For sentences providing a low degree of predictability (PL), the same phrases as those used by Kalikow et al. were used, for example, "John spoke to Mary about the beam." High context sentences (PH) were constructed such that each sentence contained at least two semantically related words leading into the target word at the end, for example, "The gymnast balanced on the beam."

A talker of standard midwest dialect was recruited to record the stimulus list. Audio recordings were made in a sound-isolated chamber, with the trained talker approximately six inches from a low-noise microphone monitored by portable sound-level meter. The sentences were spoken at a comfortable level such that the meter peaked at the same intensity for all stimuli.

Speech stimuli were specified in terms of peak and RMS levels as defined by a digital adaptation of the original Dunn & White (1947) technique (see Bell et al., 1989). The speech and noise were individually analyzed in 1/8 second epochs that were accumulated to form a distribution of short-duration window RMS values in each of 15 1/3-octave bands from 200 to 6000 Hz.

A noise was shaped to conform to the 99th percentile peak distribution of the speech stimuli in each 1/3-octave band using the prototypic data of Cox et al. (1988) to specify average peak-to-RMS values.

Insert Figure 1 about here

The problem of frequency-intensity relations changing with presentation level was overcome by using noise background stimuli which have the same general shape as the stimulus materials. With this convention, all frequencies were passed equally across the audible spectrum (see Studebaker et al., 1987). Since the noise was sufficiently above threshold levels (even for hearing-impaired), all subjects received exactly the same acoustic stimulus in sensation level in each 1/3 octave band. Thus, the significance of the shaped noise background is a critical point of protocol.

The speech-to-noise ratio for the present experiment was fixed at -2 dB, with the overall long-term RMS of the speech signal calibrated at 85 dB SPL. These absolute presentation levels were selected to ensure that the speech signal was audible for all subjects, regardless of hearing losses. The relative level of the speech and noise (-2 dB) was selected on the basis of earlier studies with the intent of collecting data between 20 and 80 percent on both types of sentence contexts employed in this study.

Lexical Characteristics. Each of the 100 target words were analyzed for frequency of occurrence in the language and for the number of phonetically similar words in the lexicon (neighborhood density). Frequency counts were taken from Kucera & Francis (1967). Neighborhood density was computed using a single sound substitution rule, similar to that discussed by Greenberg & Jenkins (1964) and used by Luce (1986).

Each list of 50 target words was arbitrarily categorized into four subsets based on median values within each list. The four subsets were labelled low-frequency/sparse neighborhood, low-frequency/dense neighborhood, high-frequency/sparse neighborhood, and high-frequency/dense neighborhood.

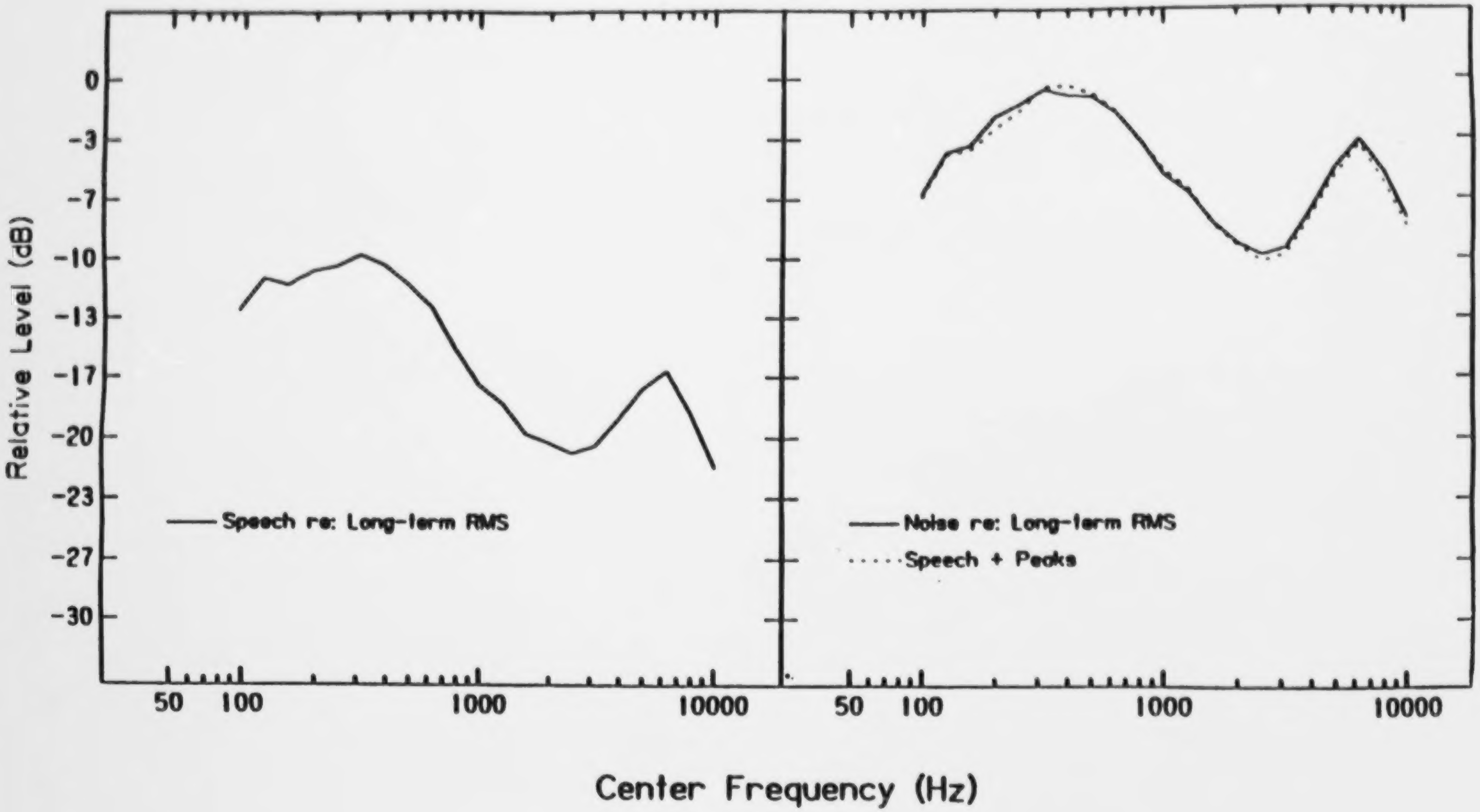


Figure 1. Left panel: Long-term RMS spectrum of the speech stimuli in one-third octave bands. Right panel: 99th percentile speech peak spectrum compared to the matched noise spectrum in one-third octave band levels re: overall RMS of the speech.

Results

A mixed four-factor analysis of variance design (2 x 2 x 2 x 2) was employed to compare differences between young and elderly listeners as a function of semantic context (low, high), relative word frequency (low, high), and neighborhood density (sparse, dense). Subject group (young versus elderly listeners) was treated as a between-subjects factor and the remaining factors were all repeated measures.

Means and standard deviations for each experimental condition are given in Table 1. As expected, all of the simple main effects were significant. The older listening group was poorer on the average at word identification [$F(1, 18) = 16.42, p < .001$]. The average number of words correctly identified overall for the young listening group was 69.6 percent compared to 39.8 percent for the elderly listeners.

Insert Table 1 about here

Target words embedded in high context sentences were more readily understood than the same words embedded in low-context sentences [$F(1, 18) = 222.14, p < .001$]. The mean percentage correct for words framed in high-context sentences was 75.6 percent, whereas in low-context sentences, percent correct was 33.7 words on average. The difference of 42 percent was constant for both age groups. Young listeners increased from 48.7 to 90.4 percent with semantic context, and elderly listeners increased from 18.6 percent to 60.9 percent. Thus, the benefit of semantic cues was constant across both subject groups.

Frequency of word usage also led to a significant difference in word identification scores. Correct identification of the target words was more likely for high-frequency words than for low-frequency words [$F(1, 18) = 48.58, p < .001$], showing a 12.5 percent advantage on the average. The young listeners increased from 62.7 to 76.5 percent, whereas the elderly listeners increased from 34.1 to 45.4 percent.

Insert Figure 2 about here

The size of the similarity neighborhood produced significant differences in percentage scores such that words from sparse similarity neighborhoods were better understood than words from dense neighborhoods [$F(1, 18) = 7.7, p < .05$]. The average percent correct for words from sparse neighborhoods was 56.9 compared to 52.4 percent for words in dense

Table 1

Means (and Standard Deviations) of Word Identification Scores for Young and Elderly Listeners as a Function of Word Frequency, Similarity Neighborhood, and Semantic Context.

			Semantic Context	
Age	Word Frequency	Similarity Neighborhood	Low	High
Young				
	Low			
		Sparse	45.5 (12.8)	84.5 (14.9)
		Dense	38.3 (12.6)	82.5 (8.2)
	High			
		Sparse	62.2 (15.4)	98.5 (3.12)
		Dense	49.2 (14.6)	96.0 (4.3)
Elderly				
	Low			
		Sparse	16.2 (13.9)	51.6 (37.8)
		Dense	17.2 (21.3)	51.6 (28.5)
	High			
		Sparse	23.6 (19.9)	73.6 (28.6)
		Dense	17.7 (14.1)	66.8 (28.3)

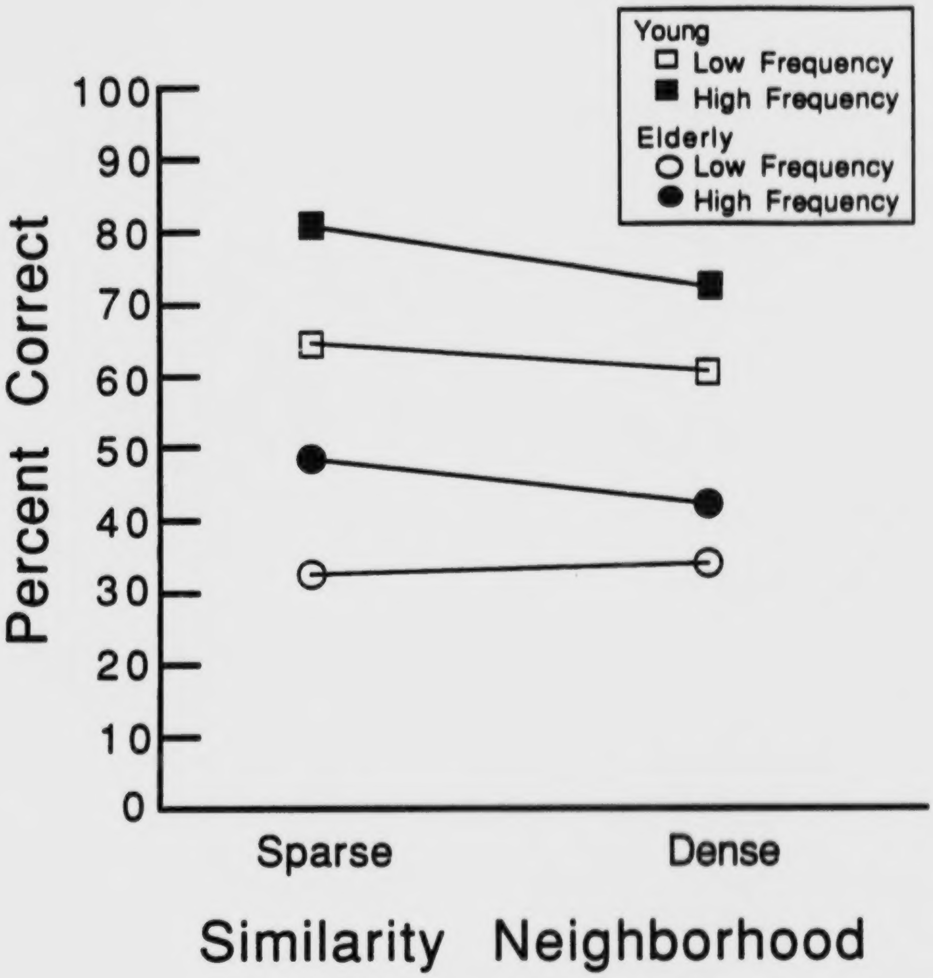


Figure 2. Percent of words correctly identified for young and elderly listeners as a function of the word frequency and similarity neighborhood.

neighborhoods, a drop of 5.5 percent in mean scores. The similarity neighborhood effect was approximately equal for young and old listeners alike, with sparse neighborhood scores of 72.6 and 66.5 percent and dense neighborhood scores of 41.2 and 38.3 percent respectively.

In addition, several two-way interactions were significant. Size of the similarity neighborhood interacted with word frequency to produce differences in the percentage of correct word identifications $[F(1,18) = 5.96, p < .05]$. As seen in Figure 2, the percentage scores for the low frequency words were approximately equal across the levels of neighborhood size, but the high frequency words had an apparent advantage in sparse neighborhoods relative to dense neighborhoods. The high-frequency advantage in sparse neighborhoods was approximately 15 percentage points (49.4 versus 61.5), but in dense neighborhoods the advantage was only 10 percent (47.4 versus 57.4).

The degree of semantic context in the carrier sentence also interacted with word frequency characteristics on percentage of correct identifications $[F(1,18) = 10.22, p < .01]$. Low-frequency words increased by approximately nine percent when presented in sentences containing semantic context relative to the same words presented in low context sentence frames. In contrast, high-frequency words increased 16 percentage points in the high context sentence condition. The mean values for low-frequency words were 29.3 and 67.6 percent for low and high context frames respectively, whereas the mean percentage scores were 38.2 and 83.7 respectively for high-frequency words (see Figure 3).

Insert Figure 3 about here

The highest order interaction present in these data involved listening group, word frequency, and semantic context $[F(1,18) = 8.00, p < .05]$. As shown in Figure 3, the younger group of listeners exhibited a constant 13 to 14 percent advantage for frequently used words over infrequently used words, regardless of the degree of semantic context available. The elderly listeners showed approximately an 18 percent advantage for high-frequency words over low-frequency words in high context sentence frames, although, in low-context sentence frames, the advantage was only 4 percent.

Discussion

These data support the Neighborhood Activation Model in several important respects, including interactive changes in word identification scores as a function of word frequency and neighborhood similarity effects. While the average frequency of the neighbors was not controlled, the size of the neighborhood relative to the frequency of the stimulus target word

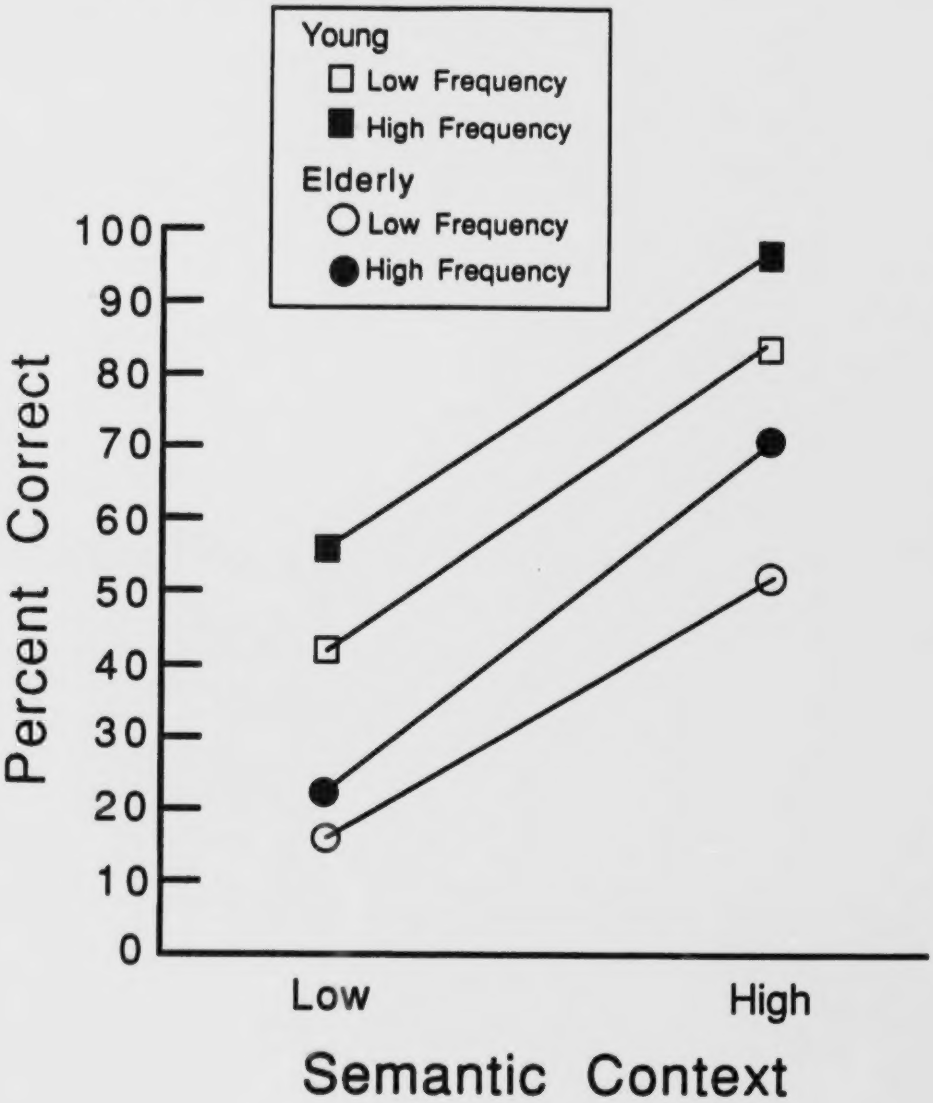


Figure 3. Percent of words correctly identified by young and elderly listeners as a function of word frequency and semantic context.

exhibited significant influence on word identification results. Assuming that large neighborhoods contain more high-frequency members on the average than sparse neighborhoods, an interactive pattern of identification results would be predicted, where high-frequency words were better understood when there were fewer phonetically similar words to interfere with the lexical decision process.

When the stimulus word usage frequency was relatively high, there was a facilitory effect on the probability of correctly identifying it. Conversely, low-frequency words were less likely to be correctly identified. The overall percentage scores for the normal-hearing young adults was typically above 50 percent. However, for these young listeners the size of the advantage of high-frequency words over low-frequency words diminished with increasing neighborhood density. But the size of the high-frequency advantage did not change with semantic context for these same listeners. Thus, even though the number of possible alternatives was reduced in the high semantic context condition, the frequency bias remained constant. Semantic activation had no effect on word frequency and lexical density effects, although it did have a dramatic effect on the overall probability of correctly identifying the target word.

The presence of word frequency effects was not surprising, although, the interaction of this variable with semantic context and age was a unique finding. The results of the present study indicated that the elderly subjects could not, or did not, use word frequency in the absence of semantic context, whereas the younger listeners exhibited word frequency effects regardless of the amount of semantic priming. If frequency effects are associated with lower order processes, then semantic activation should show no effect of the biasing caused by word usage frequency, or at least the effects should be uniform, regardless of the degree of semantic context.

Only in the presence of semantic context did word frequency effects emerge as a biasing factor for the elderly listeners, and the difference in the frequency effect for the age groups was dependent on the degree of semantic context available. In other words, the lower-order acoustic-phonetic processes were uniformly influential for the younger listeners in both contexts, but they were less influential in the lexical processes of the elderly listeners. These results demonstrated that word frequency biases occurred at a later stage than acoustic-phonetic pattern recognition. Assuming that the degraded periphery and noise conditions were particularly detrimental to the performance of the elderly group, the information was supplied primarily from the context rather than the acoustic-phonetic analysis. Since performance was poorer overall when compared to the younger group (as in Feier & Gerstman, 1980), and particularly poor in the absence of semantic cues, frequency effects cannot originate from base activation levels (see also Luce, 1986; Goldinger et al., 1989)

A tentative explanation of these results could hinge on the nature of the set of alternatives available to the listener in the lexical decision process. The elderly listening group may have tended to choose only high-frequency words from a list that contains entries based predominantly on the semantic cues available to them. The fact that the elderly listening group showed no advantage for high-frequency words in the low semantic context condition indicates that the acoustic-phonetic patterns were not the basis for the lexical decision process.

Studies of the role of context in processing continuous speech messages typically have employed distorted or masked speech signals (e.g., Miller, Heisse & Lichten, 1951; Miller & Isard, 1963), partly due the steep performance-intensity functions associated with sentence materials. Typically, the range over which intelligibility rises for highly contextual speech from near zero to optimal is on the order of 8 to 12 dB, depending on specific task parameters, noise characteristics, and threshold characteristics. In contrast, monosyllabic words or nonsense syllables rise over approximately a 20, 30, or even 40 dB range (e.g. ANSI S3.5-1969). The present results replicate this finding. At identical signal-to-noise ratios the low context sentences exhibited scores that were 37 percent lower on the average than the same words in high context sentences. These results are consistent with the gating paradigm studies cited earlier (e.g. Salasoo & Pisoni, 1985) that demonstrated that much less information is needed to identify words in the presence of semantic context than would be needed to identify the words in isolation or in sentences with only limited contextual information.

In the present experiment the background noise conformed to the long-term RMS spectrum of the speech in 1/3 octave bands. Had another spectral shaping been used, one might expect that the younger group of subjects would exhibit results similar to those of the elderly listeners. As the stimulus became more degraded, the use of acoustic-phonetic information would diminish, and words in high context sentences would show an increase in frequency bias, whereas words in low context frames would show a reduction in frequency bias in terms of identification performance. Studies have shown that one specific factor that differentially affects older versus younger persons is background noise. Older adults experience more difficulty with speech in noise (Smith & Prather, 1976; Plomp & Mimpen, 1979; Duquesnoy, 1983) than do younger individuals.

The reduced scores of the elderly listeners may have resulted from the presence of noise. In the present study the background noise was specifically shaped to match the long-term spectrum of the speech stimulus. Thus, on the average, the speech was not spectrally impoverished, rather, it was presented in an interfering background that allowed the relative spectral and temporal cues present in the speech to be passed across a broad band of frequencies. Under these conditions, the present data show that frequency and neighborhood effects were present. In the case of a noise background that deviates from the particular speech spectrum, for example a white noise (as is typically employed in speech experiments) or an interfering voice, one might expect neighborhood effects to diminish. Similarly, had the experiment employed a quiet listening environment, and used presentation levels that produced less than optimal performance levels (20 - 80 percent as in the present experiment), the outcome again may have been different. In that case, the low performance levels would have likely been the result of differentially passing frequency cues as parts of the spectrum fell below threshold. Thus, it is reasonable to assume that the noise employed in this study provided interference rather than masking that would cause a loss of spectral and temporal cues. The noise may have interfered with processing for the elderly listeners moreso than for the younger listeners; this characteristic, associated with advancing age, could have interacted with lexical and semantic processing of words in sentences.

In summary, these data have shown that word frequency effects were evident even for words presented in noise, which resulted in considerably less than optimal levels of identification performance. Previous experiments have focused on word recognition and identification in optimal environments; the Neighborhood Activation Model appears to predict performance characteristics of semantically related speech in noise as well. These results also point to a model where the elderly listener relies heavily on semantic cues in forming a set of potential candidates during the lexical decision process, possibly due to a lack of stimulus specificity either from a deteriorating periphery or from a general slowing of the nervous system (Poon & Fozard, 1980; Wingfield et al., 1985).

Further studies are required to elaborate the interaction of acoustic-phonetic, lexical and semantic processes among elderly listeners. Stimuli should be selected on the basis of their lexical characteristics and test conditions should include a finer gradient of conditions involving variables such as signal-to-noise ratio, spectral character of the interfering noise, age, audiometric configuration, linguistic competence, and semantic as well as syntactic characteristics of the context in which the words are presented.

References

American National Standards Institute (1970). American national standard method for the calculation of the Articulation Index. ANSI S3.5-1969. New York: American National Standards Institute.

American National Standards Institute (1970). Specification for audiometers. ANSI S3.6-1969. New York: American National Standards Institute.

Bell, T.S., Dirks, D.D. & Trine, T. (1989). Articulation Index importance functions for contextual speech material. *Journal of the Acoustical Society of America*, **86**, S80.

Bergmann, M., Blumenfeld, V.G. & Levitt H (1976). Age related decrement in hearing for speech. Sampling and longitudinal studies. *Journal of Gerontology*, **31**(5), 533-8.

Bosatra, A. & Russolo, M. (1982). Comparison between central tonal tests and central speech tests in elderly subjects. *Audiology*, **21**(4), 334-41.

Butler, R.N. & Gastel, B. (1979). Hearing and age: Research challenges and the National Institute on Aging. *Ann Otol Rhino Laryngol*, **88**,(5 pt 1), 676-83.

Cervellera, G. & Quaranta, A. (1982). Audiologic findings in presbycusis. *Journal of Auditory Research*, **22**(3), 161-71.

Cohen, G. & Faulkner, D. (1983). Word recognition: Age differences in contextual facilitation effects. *British Journal of Psychology*, **74**(pt 2), 239-51.

Cotton, S. & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics*, **35**, 41-48.

Cox, R., Matesich, J. & Moore, J. (1988). Distribution of short-term RMS levels for conversational speech. *Journal of the Acoustical Society of America*, **84**(3), 1100-4.

Dunn, H.K. & White, S.D. (1940). Statistical measurements on conversational speech. *Journal of the Acoustical Society of America*, **11**, 278-288.

Duquesnoy, A.J. (1983). The intelligibility of sentences in quiet and in noise in aged listeners. *Journal of the Acoustical Society of America*, **74**(4), 1136-44.

Feier, C. & Gerstman, L. (1980). Sentence comprehension abilities throughout the adult lifespan. *Journal of Gerontology*, **35**, 722-8.

Ford, J.M. & Roth, W.T. (1977). Do cognitive abilities decline with age? *Geriatrics*, **32**(9), 59-62.

Goldinger, S., Luce, P., & Pisoni, D. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-18.

Greenberg, J., & Jenkins, J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, **20**, 157-177.

Grosjean, F. (1980). Spoken word recognition processes and the gating paridigm. *Perception & Psychophysics*, **28**, 267-283.

Grosjean, F. (1985). The recognition of words after their acoustic offsets: evidence and implications. *Perception & Psychophysics*, **38**, 299-310.

Hayes, D. & Jerger, J. (1979). Low-frequency hearing loss in presbycusis. A central interpretation. *Archives of Otolaryngology*, **105**(1), 9-12.

Jerger, J., Jerger, S., Oliver, T. & Pirozzolo, F. (1989). Speech understanding and the elderly. *Ear and Hearing*, **10**(2), 79–89.

Jokinen, K. (1973). Presbycusis – masking of speech. *Acta Oto-Laryngology*, **76**, 426-430.

Kalikow, D., Stevens, K. & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled predictability. *Journal of the Acoustical Society of America*, **61**, 1337–51.

Kucera, F. & Francis, W. (1967). *Computational analysis of present day English*. Providence, Rhode Island, Brown University.

Lehiste, I. & Peterson, G.E. (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America*, **31**(3), 280-6.

Lorsbach, T.C. & Simpson, G.B. (1984). Age differences in the rate of processing in short term memory. *Journal of Gerontology*, **39**(30), 315-21.

Lowell, S.H. & Paparella, M.M. (1977). Presbycusis: what is it? *Laryngoscope*, **87**(10), 1710-7.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report 6*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A., Pisoni, D.P., & Goldinger, S.D. (in press). Similarity neighborhoods of spoken words. In Altmann (Ed.), *Cognitive Representation of Speech*, MIT Press, Cambridge.

Marlsen-Wilson, W.D. & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, **2**, 217–228.

Miller, G.A., Heisse, G.A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329–335.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M.E. Wrolstal, & H. Bouma (Eds.). *Processing of Visible Language. 1*, Plenum, New York.

Pisoni, D. (1981). Some current theoretical issues in speech perception. *Cognition*, **10**, 249–259.

Pisoni, D. (1985). Speech perception: Some new directions in research and theory. *Journal of the Acoustical Society of America*, **78**(1), 381–8.

Plomp, R. & Mimpen, A.M. (1979). Speech reception threshold for sentences as a function of age and noise level. *Journal of Acoustical Society of America*, **66**(5), 1333–42.

Poon, L.W & Fozard, J.L. (1980). Age and word frequency effects in continuous recognition memory. *Journal of Gerontology*, **35**(1), 77–86.

Roedder, D. & Cole, J. (1986). Age differences in information processing. *Journal of Consumer Research*, **13**, 297–311.

Salasoo, A. & Pisoni, D.P. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, **24**, 210–231.

Smith, R.A. & Prather, W.F. (1971). Phoneme discrimination in older persons under varying S/N conditions. *Journal of Speech and Hearing Research*, **14**, 630–8.

Spilich, G.J. & Voss, J.F. (1983). Contextual effects upon text memory for young, aged-normal, and aged memory impaired individuals. *Experimental Aging Research*, **9**(1), 45–9.

Stinson, M. & Tracy, O.A. (1983). Specificity of word meaning and use of sentence context by hearing impaired adults. *Journal of Communicative Disorders*, **16**(3), 163–73.

Studebaker, G.A., Pavlovic, C.V. & Sherbecoe, R.L. (1987). A frequency importance function for continuous discourse. *Journal of the Acoustical Society of America*, **81**(3), 1130–1138.

Weinstein, B.E. & Ventry, I.M. (1983). Audiologic correlates of hearing handicap in the elderly. *Journal of Speech and Hearing Research*, **26**(1), 148–51.

Welsh, L.W., Welsh, J.J. & Healy, M.P. (1985). Central presbycusis. *Laryngoscope*, **95**(2), 128–36.

West, R.L. & Cohen, S.L. (1985). The systematic use of semantic and acoustic processing by younger and older adults. *Experimental Aging Research*, **11**(2), 81-86.

Wingfield, A., Poon, L.W., Lombardi, L. & Lowe, D. (1985). Speed of processing in normal aging: effects of speech rate, linguistic structure, and processing time. *Journal of Gerontology*, **40**(5), 579-85.

Movement Dynamics and the Nature of Errors in Tongue Twisters: An Observation and Research Proposal[1]

Stephen D. Goldinger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] Supported provided by NIH Research Grant DC-00111-13 to Indiana University, Bloomington, IN. I thank Keith Johnson, David Pisoni, Linda Smith, Esther Thelen, Geoff Bingham, Bob Port, Michael Studdert-Kennedy, and Elliot Saltzman for their interest and comments on the current manuscript and for their suggestions for continuing research.

Abstract

Errors produced during the repetition of tongue twisters have long been considered informative about natural speech errors and speech production in general. Recent considerations of phonological retrieval processes in speech production have emphasized the role of an activation-and-selection system that is prone to systematic error when similar phonemes are selected in sequence. Examples are Dell's (1986) and Kupin's (1982) models of speech production. Kupin (1982) bases his theory on observations of errors produced by subjects during the rapid repetition of tongue twisters. In formulating and proposing his theory, Kupin dismisses physical articulator motion as an important source of error in tongue twisters. This paper reconsiders the role of actual articulator motion as a contributor to the difficulty of tongue twisters. It is argued that the kinematic dynamics of tongue twister articulation develop strong attractors that engender errors. For instance, many of the tongue twisters considered appear to require conterphasic motion through a relatively constrained area of the possible articulatory space. It is suggested that the regularities observed in this preliminary analysis justify further research on the conribution of motion dynamics to the difficulty of tongue twisters.

Movement Dynamics and the Nature of Errors in Tongue Twisters: An Observation and Research Proposal

The scientific study of language production has historically provided a troublesome methodological dilemma for language researchers. For those investigators interested in studying *speech* production, including study of the precise acoustic characteristics of speech, or the changes that occur in the speech signal under varying environmental conditions, or the complex motions of the physical articulators during speech production, the spoken utterance represents a uniquely gratifying entity for examination. The spoken utterance and its articulation are among the few behavioral gestures that can be elicited from a subject, recorded in real-time, and physically measured later along many dimensions. With respect to the limited domain of language, they are perhaps the only behaviors with these concrete properties. As such, experimentation on spoken utterances produced in the laboratory may be conducted with little sacrifice of ecological validity, a problem in psychological methodology that does not receive the concern it probably should. For the linguist or psychologist who studies language, a rich and complex domain involving abstract representations and manipulations, physically measurable speech articulation and speech waveforms become attractive experimental subjects indeed.

Unfortunately for these researchers, however, language production is part of human cognition, and is therefore extremely complex and flexible. While the flexibility of cognition is inherently pleasing to Darwinian evolutionary theorists or to tourists navigating through unfamiliar cities, it is problematic for those who would study cognitive processes with scientific rigor. Because the human cognitive system performs complex functions with such apparent ease, it is not generally possible to explain cognitive processing by mere observation. For this reason, cognitive psychology has followed the examples provided by classical psychophysics and the method of limits: Data collected in cognitive psychological experiments oftentimes consists of reaction time or error measurements. For the investigator interested in the psychological processes *underlying* the planning and production of a spoken utterance rather than the physical characteristics of the utterance itself, well-formed words and phrases simply provide too much information for systematic study to proceed. Instead, speech *errors* have provided the majority of insights into the sub-processes of language production to date. Speech errors provide the unique opportunity to compare what the speech production system intended with what it actually produced, to observe how and when the failure occurred. Such opportunities can, in turn, inform us about how utterances are prepared and produced when all goes well. The study of speech errors, however, does involve its own quandary.

The dilemma associated with speech errors is that only *true* errors allow researchers to draw inferences about underlying psychological intentions, but it is difficult to elicit ecologically valid instances of speech errors in the laboratory. Accordingly, it may be difficult to determine the level of productive planning responsible for any given speech error collected in the laboratory, whether the error was the result of an incorrect semantic, lexical, or phonological entry or if the error was the result of typical articulatory failures. Many researchers (e.g.

Fromkin, 1971) have adopted the strategy of conducting impromptu "field" experiments—Fromkin has been known to carry a small, continuous-loop tape recorder with her at all times so she may capture any errant speech errors she encounters. To the degree that entire sentences and intended sentences can be recorded faithfully, such naturalistic methods may be the most informative available. Researchers who are more interested in lower-level linguistic planning, such as the retrieval of phonological components of words have adopted experimental procedures, such as the "slips" technique (see, e.g., Dell, 1986), in which errors are induced by suddenly changing one or two segments of a phrase that subjects have already spoken repeatedly. An alternative approach to the scientific study of speech errors, however, is the investigation of the errors made during the production of tongue-twisters. Tongue twisters offer a compromise with regard to ecological validity: While the errors committed during the production of tongue twisters may not be representative of speech errors in general, at least tongue twisters reliably cause errors whether they are spoken in the laboratory or not. For researchers who are primarily interested in lower-level production systems, such ecological validity may be more than satisfactory. The relation of tongue twisters to more pedestrian speech errors will be discussed in more detail below.

Tongue twisters

Tongue twisters are phrases that consistently cause errors in production (Kupin, 1982). However, not all phrases that cause errors are tongue twisters. Kupin carefully distinguishes tongue twisters from other notoriously difficult kinds of phrases: Shibboleths and alliterative phrases. Shibboleths are phrases or words that include phonetic combinations that do not occur in the speaker's native language. Alliterative phrases (e.g. *Peter Piper picked a peck of pickled peppers*) are generally longer than tongue twisters, have more complete syntax than conventional tongue twisters, and are generally easier to say than tongue twisters. The distinction of tongue twisters from "normal sentences" is a more general delineation— normal sentences have much greater variety in their phonetic constitution than tongue twisters do. Kupin asserts that normal sentences represent random selections and orders of the phonotactically circumscribed universe, whereas the phonetic constitution and ordering of tongue twisters is far more constrained. The constraints limiting phonetic variety in tongue twisters are usually made especially salient to the hapless speaker by the common requirement that the tongue twister must be repeated several times, and as rapidly as possible.

Examples of bonafide tongue twisters by Kupin's classification are:

Brad's burned bran buns
The sixth sheik's sixth sheep's sick
Black bug's blood

The present approach

The goal of this paper is to begin explorations of tongue twisters with an eye toward a minimalist explanation of the most common errors reported. The most common sorts of hypotheses offered in explanation of such errors are based on assumptions of excessive demands for articulatory mechanics (e.g. Linder, 1969, cited in Kupin, 1982), or of confusions at the motor planning stages of production (e.g. Kupin, 1982; Dell, 1984; Shattuck-Hufnagle, 1986). Certainly, most or all of the theories that have been developed to explain these phenomena are well-supported by behavioral data, and will almost certainly prove to have at least part of the story correct: At some level, motor planning and physical production *must* be intimately involved in the processes by which speakers produce utterances and by which speakers err in such productions. However, the difficulty that theorists have experienced in communicating their ideas has been the problem of distinguishing tongue twisters from ordinary phrases.

When tongue twisters are considered introspectively or experientially, there is general agreement among speakers of the language that tongue twisters are indeed qualitatively different from normal household sentences. Moreover, when one attempts to describe what it is about tongue twisters that makes them different from normal sentences, it becomes readily apparent that tongue twisters tend to consist of a small subset of speech sounds that are said repeatedly. This observation then leads to the "intuitive" approach to explanation—the temptation to introspect on how *we* would contend with such a phrase and to conclude from our introspections that "yes, indeed, this is hard to say because of these repeated and/or similar phonemes." Unfortunately, it would be very difficult to use these notions to *predict* a priori that any given sentence would be a tongue twister without resorting to such self-experimentation. Aside from these "intuitive" differences, it remains unclear why tongue twisters are so difficult and normal phrases (and especially alliterative phrases) are so easy. Put another way, if one could witness *only* the chaotic acrobatics that the speech articulators and planning system must perform during casual conversation, it is not clear that a "difficult" series of motions could ever be reliably distinguished from more "simple" ones. In fact, my *own* introspections tell me that if I had no knowledge of tongue twisters or their phonetic characteristics, I would most likely predict that the utterances that contain the *greatest* variability would be the most difficult, contrary to the established phenomenon. It is with regard to this problem that a dynamical description of spoken phrases may offer unique insights.

The work described in this paper represents a preliminary attempt to perform the cumbersome task alluded to above– to witness the articulatory acrobatics associated with tongue twisters and other phrases and try to draw some general inferences about how such phrases may fundamentally differ. A case will be made for considering tongue twisters from a more general stance from kinematics and movement dynamics, rather than from phonological or activation-based theories of production. In witnessing these motions, however, we will not simply peer into an open mouth, as if we were using x-ray photography. Instead, we will observe a more abstract state space that is derived from the mouth but is less complex. The space is defined by three continuous scales determined by articulatory gestures that occur during speech. Figure 1 shows the state space of all possible values that will circumscribe the articulatory possibilities considered here:

Insert Figure 1 about here

The X-axis of the state space represents the position in the mouth of the tongue body, a good indicator of consonant quality. The scale is continuous from completely back (as in a glottalized /g/ sound), to completely front (as in a dentalized /t/). The Y-axis represents degree of lip rounding, a good indicator of vowel quality. The scale is continuous from completely unrounded (as in /i/, in "deed"), to completely rounded (as in an exaggerated /u/, as in "food"). The Z-axis represents tongue body height, a good indicator of both consonant and vowel quality. The scale is continuous from high (as in the consonant /k/) to low (as in retroflex /r/). Three-dimensional represention of phonetic space using these same articulatory dimensions have been adopted for various reasons, such as describing the vowel space, by Lisker (1988), Terbeek (1977) and others.

In the approach adopted here, the representation of speech within this state space takes the form of continuous trajectories that begin with the initial gestures of the utterance and "move" through all subsequent gestures until the utterance is complete. Each point along the trajectory is interpreted as a point in three-dimensional space corresponding to a triangulation of the values of all three dimensions. Thus, although the representations adopted are abstractions from real mouths, all the estimations are based on real physical dimensions. To faciltitate understanding of these figures, most trajectories represent fairly short stretches of speech, not exceeding two syllables. *It is essential* to note that all illustrations of articulator positions presented here are *only* estimations. The actual values of these motions along these three scales could show considerable variability. Nevertheless, the approximations offered here should be relatively close to "prototypical" productions of these utterances, and the general conclusions that are drawn from these approximations should not be affected by any minor alterations.

A final point about the representations of tongue twisters provided in this paper: Although coarticulatory influences would certainly yield wide variations in actual productions of the phrases considered below, for the purposes of the estimations considered in this paper, a general constraint was imposed. It was assumed that several of the consonants could be articulated in widely varying locations of the state space. For example, tongue height and position are not closely specified for bilabial consonants such as /b/ or /p/. Accordingly, the estimated postions for such consonants were "moved" from a neutral position in the space

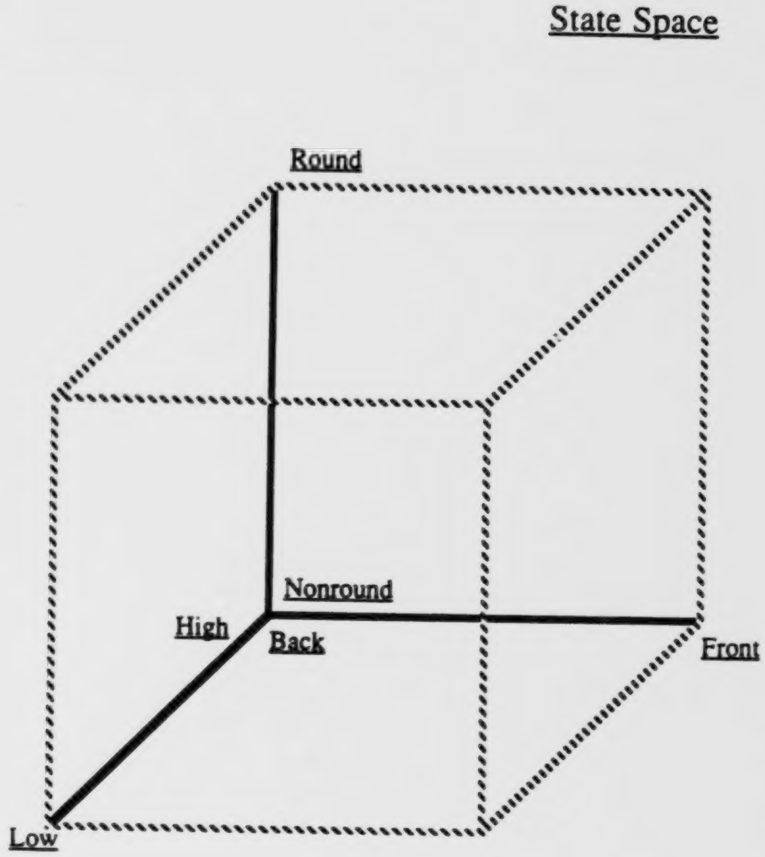

State Space
Round
Nonround
High
Back
Front
Low

Figure 1. A state space for the desciption of speech articulation.

in the directions of the positions of the vowels that followed the consonants. Given this strategy for determining consonant positions in the state space, it was necessary that vowel estimations follow a predetermined set of coordinates that do not change from one example to the next. It was *not* assumed that the exact same vowel coordinates are reached in all phrases, but it was assumed that the trajectories described by the tongue twisters *approach* vowel coordinates that are invariant across examples. (In actual productions, the degree of approximation to "canonical" vowels may be influenced by many factors, including degrees of stress on the syllables and speaking rates.)

The estimations of vowel coordinates in the state space are shown in Figure 2.

Insert Figure 2 about here

Representations of tongue twisters and other phrases

The objective having been stated, this section of the paper begins examinination of representations of tongue twisters and other phrases. Three different kinds of phrases are examined: Kupin (1982) describes two different kinds of tongue twisters, based on the kinds of errors they typically induce. The first kind of tongue twister induces what he refers to as "phonetic errors," errors in which the speaker tends to "trip" over a repeated consonant or consonant cluster. A common example is the /bl/ in the tongue twister *black bug's blood*. The second kind of tongue twister induces what Kupin refers to as "well-formed errors," which are based on rhythmic perseverations or substitutions. A common example is the tongue twister *pure food for pure mules*. These two kinds of tongue twisters are the first two kinds of phrases considered here. The third kind of phrase examined is the "normal phrase" that does not cause errors for most speakers, even when the phrase is repeated rapidly.

Proper interpretation of all the figures in this paper depend on not only consideration of the trajectories displayed, but on several critical inferences as well. The information provided in the figures is the location of the articulators at different points in time, and their direction of motion through space. However, for the sake of clarity in the displays, not all articulatory motion is considered. Specifically, each trajectory represents one or two syllables as *discrete* events. Obviously, a full representation of these tongue twisters would include connections from the end of one syllable to the beginning of the next. Similarly, recall that tongue twisters are assumed to be spoken repeatedly. As such, it would be appropriate to represent articulator motions from the completion of one phrase to the initiation of the next. Unfortunately, such enriched displays are difficult to negotiate, and the communicative purpose of the figures would be sacrificed.

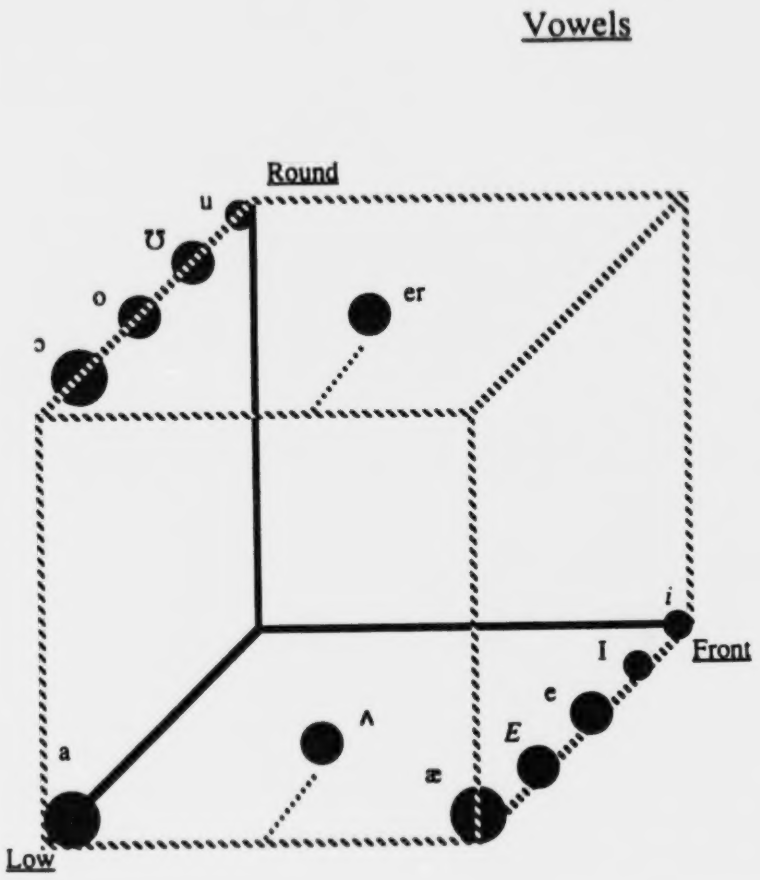

Vowels
Round
u
ʊ
o
ɔ
er
i
I
Front
e
E
ʌ
a
æ
Low

Figure 2. The locations of the vowels of American English in the state space. From Terbeek (1977).

The burden is therefore shifted to the reader to infer the undrawn lines for each figure. In all the examples considered in this paper, the articulators are assumed to move in a smooth motion from the termination of one trajectory to the beginning of the next. As such, it should be easy for the reader to infer the missing information and synthesize a complete, repetitive cycle through the space. Consideration of these unspoken portions are important to adequately portray the full course of motion through the state space for each tongue twister, and therefore to appreciate the claims made about each example.

Considering our first tongue twister, Figure 3 shows an abstract representation of the "phoneme-based" tongue twister *Peggy Babcock*. The unfilled trajectory denotes the segment " Peggy," the striated trajectory denotes the syllable "Bab," and the cross-hatched trajectory denotes the syllable "cock." The trajectories follow the direction of the arrowheads drawn along their lengths.

Insert Figure 3 about here

Two important aspects of this pattern should be noticed. First of all, it should be noted that although there are clearly redundant segments in this tongue twister, there are not an exceptional number of them. Also, the overall patterns of the trajectories for "Peggy" and "Babcock" are somewhat dissimilar; they diverge for a substantial portion of their total lengths. It is clear from this example that a strongly constrained phonetic content is not neccessary for an utterance to be a tongue twister, as shall be seen in several more examples below. Rather than the specific segments composing *Peggy Babcock*, it appears that the specific locations in the state space and the order in which they must be approached are the determinants of difficulty.

It has been noted by Kupin (1982) and others that typical errors produced in tongue twisters often involve sequencing errors. For instance, the initial consonants in *she sells sea shells* describe an ABBA pattern, whereas speakers tend to prefer ABAB patterns. However, Kupin makes an argument that these sorts of errors (indeed, all tongue twister errors) represent failures of a phonetic selection mechanism that retrieves phonetic features from the lexicon during production planning. The basis for errors in Kupin's two stage priming-plus-search model is very similar to the basis for errors in Dell's (1986) model of speech production, depending on the concept of activation among similar segments as a basis for faulty selections. In fact, Kupin strongly asserts that alternating articulatory movements make little or no contribution to the difficulty of tongue twisters:

> Tongue twisters do not literally twist the articulators. That is, they do not depend for their effect on manipulations that are physically difficult. Consider,

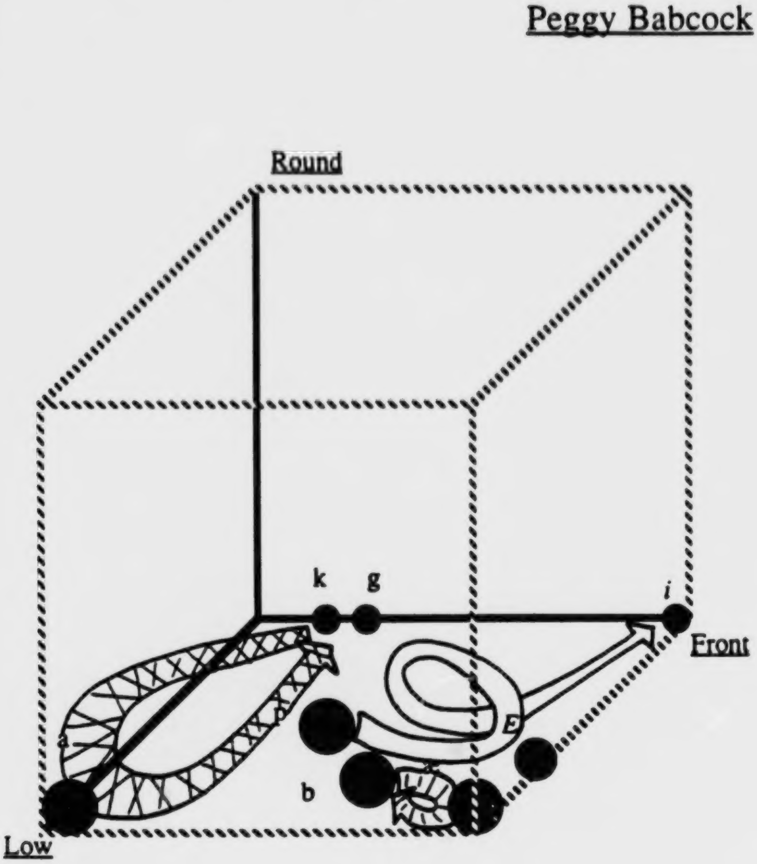

Peggy Babcock
Round
k
g
i
Front
a
b
Low

Figure 3. Articulatory trajectories of *Peggy Babcock*. Unfilled trajectory is *Peggy*, single hatched trajectory is *Bab* and double hatched trajectory is *cock*.

> for example, how excruciatingly slowly one must articulate in order to avoid making an error on *this zither*. This tongue twister is one that characteristically leads to the intuition that one simply does not have time to accurately move one's articulators from one configuration to the next. I claim that this intuition is wrong and that one has more than enough movement time. (Kupin, 1982, p. 9).

The present paper is an attempt to re-consider the speaker's articulatory intuitions that Kupin dismisses in this quote. Models of speech production based on notions of spreading activation and selection mechanisms are very powerful and are well-founded in the perceptual literature. Nevertheless, when the articulation of tongue twisters is considered from an alternate perspective, the hypothesis that physical motion may engender errors appears plausible once again. Returning to the example shown in Figure 3, one can easily recognize that "sequencing" problems like those of *she sells sea shells* are involved in *Peggy Babcock* as well. However, whereas in *she sells sea shells* the ABBA pattern is directly related to the actual initial segments of the words, in *Peggy Babcock* the pattern ABAABB is described, in which the As are front-articulated consonants, and the Bs are back-articulated consonants. The main thesis of this paper is that these sorts of sequences are not hard because the constituent segments are highly primed and prepared for faulty selections, but because the movement dynamics associated with their articulation contain mutually inhibitive attractive states.

It is well-known that any physically moving system can be described in terms of attractive states (see, e.g., Abraham, 1989). Principles of dynamic systems are easily applied to kinematic systems of simple or complex oscillators, such as the articulators defining the state spaces used in the present paper. Kelso and his associates (Kelso, Saltzman, & Tuller, 1986; Schöner & Kelso, 1988; see also Saltzman & Munhall, 1989) have discussed the dynamics of speech production at length. One of the main findings that Kelso has observed is that kinematic systems are extremely sensitive to phase timing relations among components. Accordingly, it is is difficult for a person to swing his or her hands back and forth with the two hands moving out of phase with each other; the hands rapidly assume an in-phase movement pattern. Phase relations in movement have also been widely discussed in the context of speech (e.g., Kelso et al., 1986; Saltzman & Munhall, 1989; Lubker, 1986). With respect to tongue twisters, it is clear that the ABAABB pattern of *Peggy Babcock* is counter-phasic whereas the common error "Pebby Babpop" (AAAAAA) is clearly phasic. Tracing the overall trajectory described by *Peggy Babcock* makes the phase shift especially salient. It is this sort of counter-phase relation that seems to be general to many tongue twisters. The nature of kinematic dynamics requires that the system degenerate to a uniform phase pattern, especially if the system is under loose control of feedback, as in rapid speaking. The specific size and locale of the attractors in different tongue twisters may be variable, and the pertubations that arise from them may show some differences, as is dicussed below, but they all originate from articulatory movement dynamics.

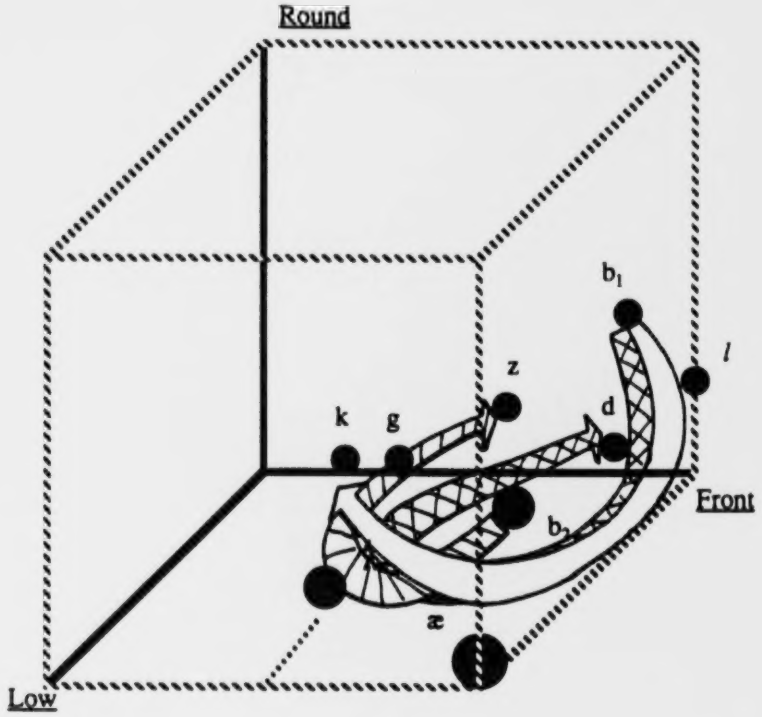
One final point should be emphasized before considering further examples. The case argued in this paper is for the viability of an articulatory basis for the difficulty of tongue twisters. I hope to demonstrate that the approach is sufficient to explain the phenomena as well as more abstract planning models, such as Kupin's model. Unfortunately, as this work is still rather preliminary, with no empirical results available yet, I can not demonstrate that a dynamic systems approach to the explanation of tongue twisters is *superior* to higher-level models. This is especially true of Kupin's model, which is largely based on notions of phonetic similarity that are based partially on articulatory properties of different phonemes. It is important that the limitations of the present work are respected. Therefore, the remaining examples presented in this section of the paper should be considered *not* as strong evidence against other models, but merely as pedagogical vehicles to communicate the generality of an articulatory dynamical approach.

Given the motivations and caveats enumerated above, consider next Figure 4, which displays the "phoneme-based" tongue twister *black bug's blood.* The unfilled trajectory represents "black," the striated trajectory represents "bug's," and the cross-hatched trajectory represents "blood." As in Figure 3, the trajectories all follow the arrowheads drawn along their lengths.

Insert Figure 4 about here

Again, as in *Peggy Babcock*, two key aspects of these patterns emerge: The overall patterns are somewhat similar, but do differ significantly. Nevertheless, there are certain points that are very close to each other in articulatory space that must be encountered repeatedly, and in a counter-phasic pattern. The key distinction in this tongue twister is /b/ and /bl/, which must be alternated by the speaker. When the tongue twister is spoken repeatedly, these two points must be reached in the alternating pattern: /bl/–/b/–/bl/–/bl/–/b/–/bl/... etc., describing a counter-phasic ABAABAABA sequence. Borrowing language from dynamical systems literature, in repetitions of both *Peggy Babcock* and *black bug's blood*, it seems that *point* or *local* attractors are engendered. To borrow terminology from Abraham (1989), if we consider the plane described by these trajectories as a landscape in space, *local minima* may be expected to develop wherever the trajectories pass a given point. Such minima, or "wells," may grow deeper with each successive pass over them, and therefore become stronger and stronger attractors. As the counter-phasic motion through the state space motivates the system to settle into a phasic pattern, the strongest attractor should prove an accurate predictor of the eventual error produced. For example, in *black bug's blood,* the cluster /bl/ must be uttered twice for every instance of the neighboring consonant /b/. Following the notions of local minima and point attractors, we would expect that the trajectories for /b/

Black Bug's Blood



Round
b_1
l
z
k
g
d
Front
b_2
æ
Low

Figure 4. Articulatory trajectories of *Black Bug's Blood*. Unfilled trajectory is *black*, single hatched trajectory is *bug's* and double hatched trajectory is *blood*.

would tend to be attracted to the point for /bl/ instead. Indeed, "black blug's blood" is the error most speakers report experiencing when trying to repeat this tongue twister several times (Kupin, 1982).

We have now considered two examples of "phoneme-based" tongue twisters, and will next examine several "pattern-based" tongue twisters. These examples demonstrate that there are noticeable differences in the nature of the attractive states invoked by the tongue twisters, but the dynamics underlying the difficulty in production is unchanged. Figure 5 shows an abstract representation of the "pattern-based" tongue twister *toy boat.* The unfilled trajectory denotes the segment "toy," and the striated trajectory denotes the segment "boat."

Insert Figure 5 about here

As in the previous examples, it is readily apparent that the trajectories for "toy" and "boat" require a counterphasic motion (indeed, a virtual reversal) through the state space. This again represents a phase shift in the middle of the utterance, but the shift is global across a larger stretch of the trajectories. To return once again to speculations derived from dynamical systems theory, it could be argued that the trajectories described during the production of such "pattern-based" tongue twisters can develop *cyclic attractors.* In more general terms, it is reasonable to assume that if the articulatory system is required negotiate opposite pathways through a rather constrained area of the possible phonotactic space over and over again, both of the syllables' cycles will assume the properties of attractors. If the system is stressed by rapid repetition of the phrase, then one of the cycles will become the stronger attractor and will subsume the other. Another possibility is that some hybrid cycle representing an "average" of the two will become the attractive minimum cycle. If this were the case, the commonly reported errors of "toy boyt", "toe boat," or something decidedly incomprehensible would be expected.

Thus, we have seen some indication that the two different kinds of tongue twisters that Kupin identifies by error type may also be distinguishable by descriptions of their dynamic attractive states. I have argued that all the errors reported in tongue twisters may be accounted for by simple articulatory dynamics, rather than by a more complex model of activation and selection of phonological entries. Before moving on to "normal phrases," however, we examine one last example of a "pattern-based" tongue twister, along with its most common mispronunciation. The upper panel of Figure 6 shows the tongue twister *unique New York.* The unfilled trajectory denotes the segments for "unique," whereas the striated trajectory denotes the segments for "New York." The lower panel of Figure 6 shows the common mispronunciation for this tongue twister, *unique Yew Nork.* Again, the unfilled

Toy Boat

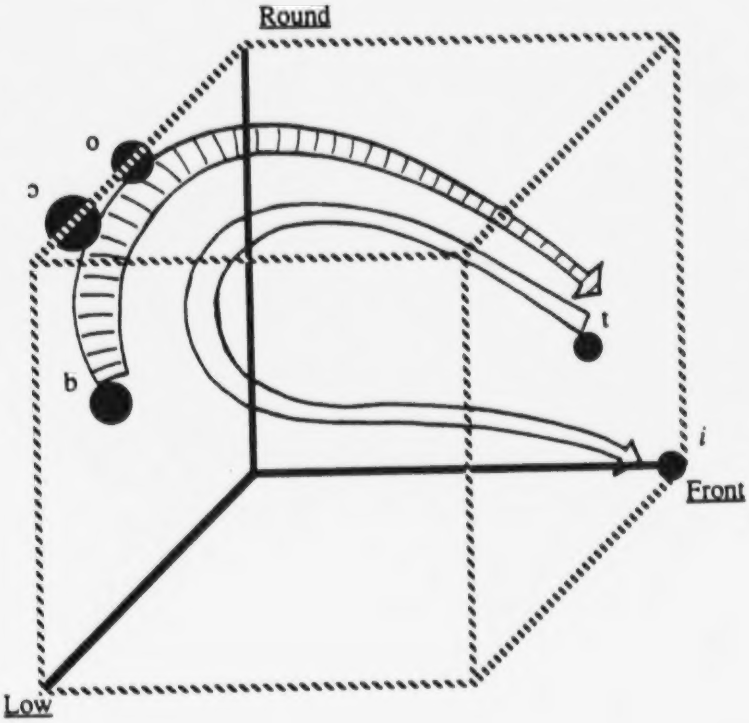

Round
o
ɔ
t
b
i
Front
Low

Figure 5. Articulatory trajectories of em Toy Boat. Unfilled trajectory is *toy* and hatched trajectory is *boat*.

trajectory denotes the segments for "unique," whereas the striated trajectory denotes the segments for "Yew Nork."

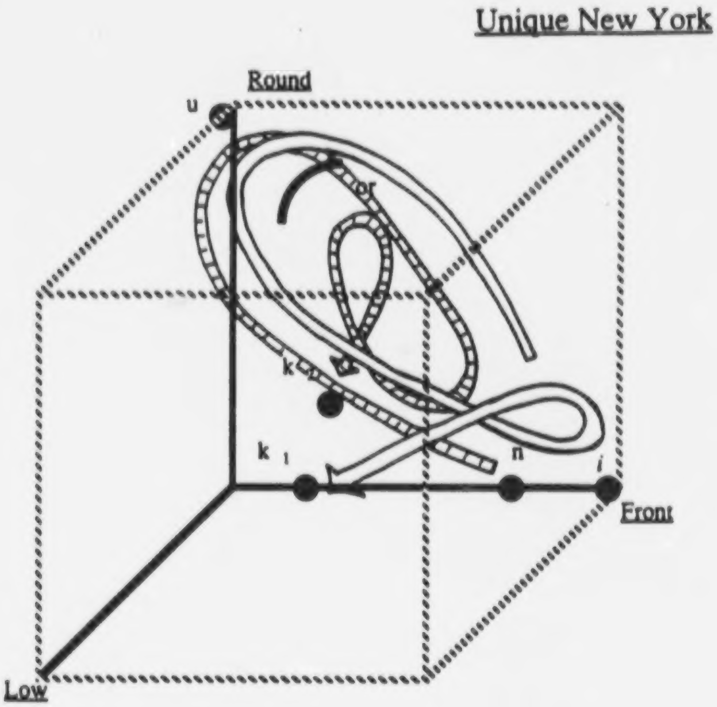
Insert Figure 6 about here

As with the tongue twisters, *toy boat*, *unique New York* consists of two trajectories that follow closely neighboring paths through the articulatory space defined here, but the trajectories pass through these areas while heading in opposite directions. Once again, the cycle for "unique" is out of phase with the cycle for "New York." Work on dynamical physical systems done by Schöner and Kelso (1988) has shown that cyclic patterns are strongly attracted to phase-consistent rhythms. Given this observation, if the difficult nature of tongue twisters such as *unique New York* are somehow originated by competitive cyclic attractors, we would expect the counter-phasic sequence *unique New York* (ABBA) to be erroneously replaced by *unique yew nork* (ABAB). This is indeed the most common error, as Kupin (1982) reports and, as can be seen in the lower panel of Figure 6, the errorful utterance consists of two phase-consistent cycles. As one might expect, *unique yew nork* is considerably easier to repeat several times than *unique New York* is.

We have now examined several different examples of tongue twisters and have noted that dynamical systems interpretations of these representations are consistent with the difficulties speakers report. Before this preliminary investigation can be closed, however, it is necessary to examine some examples of "normal utterances," phrases that do not cause speakers any profound difficulty during repeated productions. Figure 7 shows the abstract representation of the phrase *Golden Gate Bridge*. The unfilled trajectory represents "golden," the striated trajectory respresents "gate," and the cross-hatched trajectory represents "bridge."

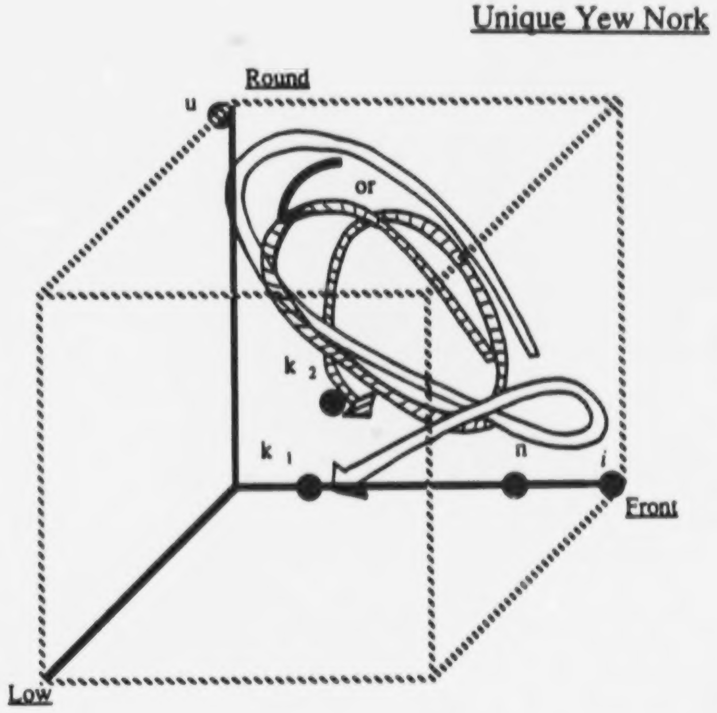
Insert Figure 7 about here

As Figure 7 shows, there is a great deal of variety in the pathways approaching the segments necessary to produce the utterance *Golden Gate Bridge*. Indeed, the trajectories described by this utterance through the space are quite distinct, appearing nearly random. Returning to dynamical systems terms, if attractors originate from repetitive points or cycles, this utterance does not seem inclined to engender any such attractors. Of course, throughout the preceding discussions of attractors and errors produced in tongue twisters,



Unique New York

Round
u
or
k 2
k 1
n
i
Front
Low



Unique Yew Nork

Round
u
or
k 2
k 1
n
i
Front
Low

Figure 6. Articulatory trajectories of *Unique New York* (top panel) and *Unique Yew Nork* (bottom panel). Unfilled trajectory is *unique*. Hatched trajectory is *New York* in the top panel and *Yew Nork* in the bottom panel.

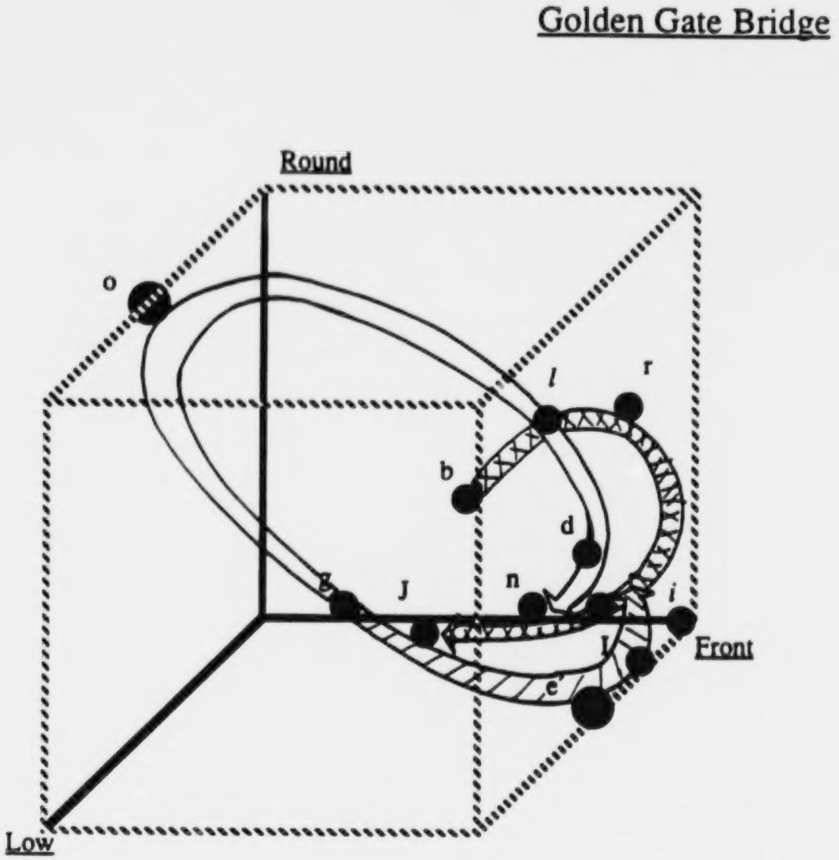

Golden Gate Bridge
Round
o
l
r
b
d
g
J
n
i
I
e
Front
Low

Figure 7. Articulatory trajectories of *Golden Gate Bridge*. Unfilled trajectory is *golden*, single hatched trajectory is *gate* and double hatched trajectory is *bridge*.

I have repeatedly mentioned the obvious control parameter of the system, which is rate of production and repetition. It appears that repetitive passes over points or cycles will only yield strong attractors if the instances of repetition are close together in time. Therefore, tongue twisters become easier if spoken more slowly, and even *Golden Gate Bridge* becomes difficult if spoken very rapidly. The important notion here, however, is that when all else is equal and all phrases are spoken at a uniform rate, "normal phrases" are much easier than tongue twisters. Accordingly, their pathways through the state space are less tightly constrained, do not require multiple phase shifts, and are replete of competitive attractors.

This notion is again illustrated in the "normal" phrase *puppy dog*, shown in Figure 8. The unfilled trajectory denotes "puppy," whereas the striated trajectory denotes "dog."

Insert Figure 8 about here

Again, as in *Golden Gate Bridge*, note that the trajectories described for *puppy dog* exploit a more variable amount of the possible state space, and there is little basis for predicting the development of any strong attractors. This is not to imply, however, that only random selections of the phonetic space can constitute "normal phrases." Tightly constrained selections of the phonetic inventory need not be experienced as tongue twisters, so long as the particular pathways connecting the phonetic gestures respect phase relations. An example of a "normal phrase" that occupies a relatively small portion of the articulatory space, but is phase-consistent is *free pizza*, shown in Figure 9:

Insert Figure 9 about here

Figure 9 shows an example of a phrase that may actually *depend* on the development of a strong cyclic attractor to make the repetitive task even easier. This final example of *free pizza* may be taken as an existence proof that neither mere phonetic inventory or tight cyclic articulations are sufficient to engender the difficulty of production indicative of tongue twisters. Principles of the movement dynamics of the articulators appear to adequately describe what makes tongue twisters more difficult than simple generic phrases.

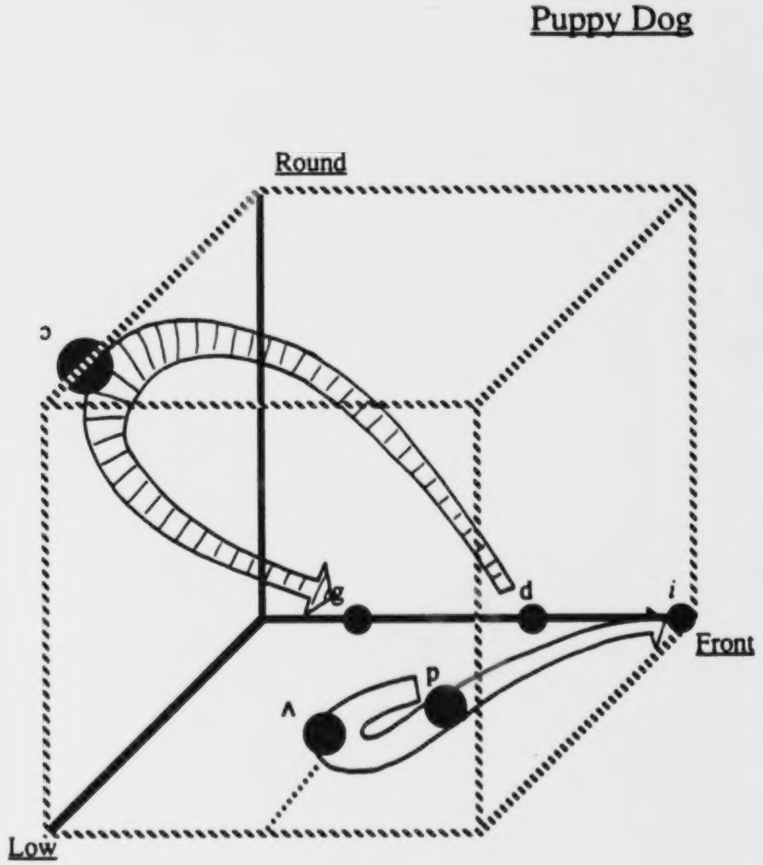

Puppy Dog
Round
ɔ
g
d
i
Front
p
ʌ
Low

Figure 8. Articulatory trajectories of *Puppy Dog*. Unfilled trajectory is *puppy* and hatched trajectory is *dog*.

Free Pizza

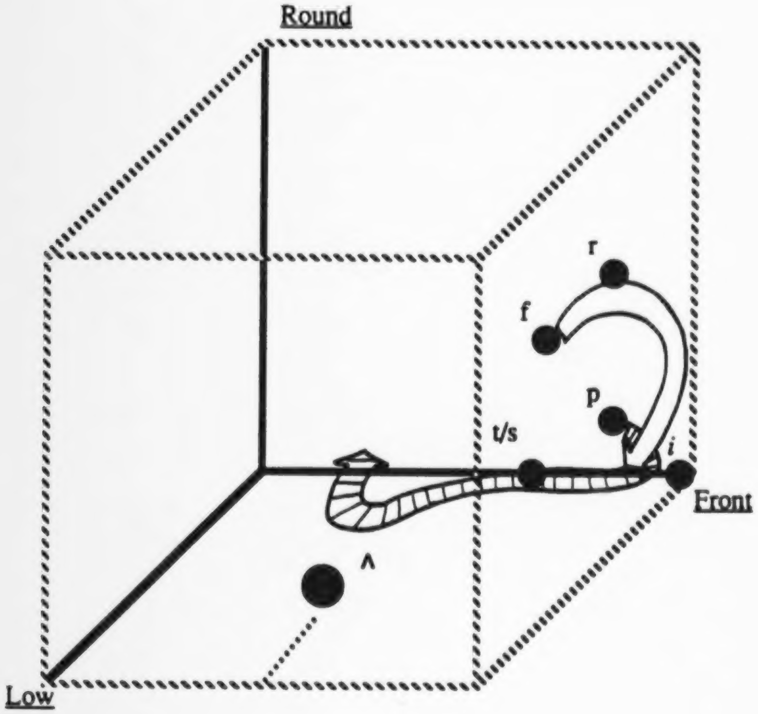

Round
r
f
p
t/s
i
Front
^
Low

Figure 9. Articulatory trajectories of *Free Pizza.* Unfilled trajectory is *free* and hatched trajectory is *pizza.*

Conclusions, caveats, and a call for further investigation

The approach that has been advocated in this preliminary paper seems to show promise in helping us understand, in a qualitative sense, what it is that makes a tongue twister a tongue twister instead of just a more pedestrian phrase. It is an approach that does not rely heavily on introspective reports from thick-tongued speakers or phonological theory, and that can be readily applied to the variety of examples one could accumulate. Principles of dynamical systems and explorations of abstract state spaces have much to offer for further investigation of complex systems such as the speech articulation system. However, further work is necessary to determine the adequacy of the account for those phenomena the approach can address, and there remain some difficult questions that the present approach may not be able to adequately address. I conclude the present introduction to these ideas by first discussing what I consider the next important steps for investigation, followed by a discussion of some more difficult issues that should be addressed in the future.

The next logical steps for investigation of the claims made here are as follows: The obvious first necessary step is actual data collection. The state spaces employed in the descriptions above were circumscribed by three dimensions that were derived directly from directions of tongue and lip movements during speech. These dimensional selections were not arbitrary. Although the present discussion was not based on empirical observations, the estimations of articulator positions were intended to be empirically evaluated. Accordingly, the dimensions were not selected to correspond to, for instance, binary feature-present / feature-absent values, which are of a more abstract, derived nature. The selection of dimensions based on attributes of real mouths makes it possible to re-create the trajectory representations used here with real data. Techniques such as x-ray pellet tracking could be easily applied to monitor tongue height, position, and lip rounding, allowing direct evaluation of the movement dynamics found in tongue twisters. (Additionally, empirical observations may allow for a more formal, quantitative presentation of the kinematic dynamics involved. Such analyses would be an important component of further research, provided that the qualitative appearance of the fit between theory and data remained intact). Although my linguistic intuitions are fair, and the estimates presented here may be reasonably close to prototypical productions of the utterances considered, the value of this approach to explanation can only be adequately assessed after production data is available.

The second logical step for further research, presuming the empirically-derived representations still displayed the intriguing properites of the estimates presented here, would be to use the observed dynamics as a sole basis for *prediction* of new tongue twisters. The analyses presented here and those proposed for pellet-tracking data are decidedly ad-hoc: Perhaps the only way to properly evaluate this approach is to determine its predictive power for utterances that are not tongue twisters by convention, as the tongue twisters considered here were. An experiment using simple measures of difficulty, such as latencies to first errors, could provide far more knowledge than any number of analyses such as those presented here.

Beyond the theoretical questions that relate directly to the claims made in this paper, a host of further questions demand further attention. This is especially true of questions that relate in one way or another to the notion of levels of linguistic planning: At what *level* do these attractors attract? At first glance, since the state spaces presented here were derived from physical articulatory gestures, it is tempting to conclude that the phenomenon of attractive states' establishment during the production of tongue twisters is a purely motor-based effect. In other words, contrary to Kupin's (1982) claim, perhaps tongue twisters twist tongues. The most extreme version of this hypothesis would totally segregate actual motor output from motor planning, leading to the hypothesis that it is *only* the actual motion that makes tongue twisters difficult to produce. However, there is evidence that is inconsistent with such a strong claim. For example, Dell (1977) and McCutchen and Perfetti (1982) has shown that speakers will report errors even during silent rehearsal of tongue twisters (the "visual tongue twister effect"). Effects such as this suggest that perhaps the true locus of these attractive states' effects during production of tongue twisters is at a motor planning stage (Kupin, 1982). The validity of this sort of phenomenon makes it far more difficult to distinguish a view based on motor dynamics from a view based on the proper sequential selection of activated features. Nevertheless, despite appearances, the "visual tongue twister effect" is not completely inconsistent with a motor dynamic view of the difficulty of tongue twisters. If one simply makes the modest assumption that lexical access[1] involves activation of several types of knowledge, including not only semantic knowledge, but orthographic, phonological and *productive* knowledge as well, it can still be argued that movement dynamics are responsible for the difficulty of silent tongue twisters. Fully distributed lexical representations of this sort have been proposed several times in the literature, perhaps most recently by Seidenberg and McClelland (1989). The concession that motion dynamics can make even silent tongue twisters difficult weakens the approach somewhat, but the interesting theoretical stance is not substantially modified. Finally, it is worth noting that, whereas an activation and selection model would predict no differences, I am reasonably sure that a motion dynamic view would predict that spoken tongue twisters should be harder than silent tongue twisters. This is, of course, a very difficult prediction to test and I know of no data that bear on the comparison directly.

Many discussions have appeared in the speech production literature about the possibilities of multiple planning stages for speech production, ranging from the most abstract level of generating the intended message from linguistic primitives to the fine-grained planning of articulator motion (see, e.g., Stemberger, 1983, for a review). At present, the work described here cannot provide unambiguous support for either side of the debate. However, if the locus of the effects is at a planning stage, it is interesting to note that the motor plans generated do not seem to be based on phoneme-by-phoneme instructions. Instead, as several theorists (e.g. Abbs, Gracco, & Cole, 1984; Browman & Goldstein, 1987) have claimed, it appears that

---

[1] The term "lexical access" is used here in the sense that it is used by Dell (1986) and other speech production researchers. It is taken to imply access to the lexicon "from the inside," rather than access to the lexicon in response to a stimulus impinging the ear.

the instructions given to the motor system are more wholistic in nature, perhaps specifying syllable-sized units. Phoneme-by-phoneme instructions should not be sensitive to cyclic attractors that span numerous segments, but the speech planning system seems to show just such sensitivities.

Finally, what can errors produced during the repetition of tongue twisters reveal about errors produced during casual speech? This is a more difficult question that again addresses the issue of multiple stages of planning. Many of the errors that speakers produce during casual speech are qualitatively different from the sorts of errors that have been addressed here. For instance, common speech errors may involve substituting an early word in a sentence with a word intended for a later position in the sentence, as in the error, *It makes the warm breather to air* for the intended sentence, *It makes the air warmer to breathe* (Garnham, 1985). Other times, words that are semantically associated with each other (e.g. opposites) will be mistakenly substituted for each other. Another important finding is that words that are substituted for each other in casual speech errors maintain their syntactic class (Fromkin, 1971). Certainly, findings such as these will require a more in-depth treatment and a level of analysis far more abstract than the tongue twisters considered here have received. The most difficult problems in speech production will certainly involve these complex domains of abstract representations and manipulations– a state space where angels fear to tread.

References

Abbs, J.H., Gracco, V.L., & Cole, K.J. (1984). Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. *Journal of Motor Behavior*, **16**, 195–231.

Abraham, R.H. (1989). *Complex Dynamical Systems.* Santa Cruz: Aerial.

Browman, C.P., & Goldstein, L. (1987). Tiers in articulatory phonology, with some implications for casual speech. *Haskins Laboratories Status Report on Speech Research*, **SR-92**. New Haven.

Dell, G. (1977). Slips of the mind. In Paradis, M. (Ed.), *The fourth Lacus forum*, Columbia, S.C.: Hornbeam Press.

Dell, G. (1984). Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 222–233.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283–321.

Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language*, **47**, 27–52.

Garnham, A. (1985). *Psycholinguistics: Central topics.* New York: Methuen.

Kelso, J.A.S., Saltzman, E.L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, **14**, 29–60.

Kupin, J.J. (1982). *Tongue twisters as a source of information about speech production.* Doctoral dissertation, Yale University. Published by Indiana University Linguistics Club.

Lubker, J. (1986). Articulatory timing and the concept of phase. *Journal of Phonetics*, **14**, 133–138.

Lisker, L. (1988). On the articulatory interpretation of vowel "quality": The dimension of rounding. *Haskins Laboratories Status Report on Speech Research*, **SR-95/96**. New Haven.

McCutchen, D., & Perfetti, C.A. (1982). The visual tongue-twister effect: Phonological activation in silent reading. *Journal of Verbal Learning and Verbal Behavior*, **21**, 672–687.

Saltzman, E.L., & Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 333–382.

Schöner, G., & Kelso, J.A.S. (1988). Dynamic pattern generation in behavioral and neural systems. *Science*, **239**, 1513–1520.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523–568.

Shattuck-Hufnagle, S. (1986). The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology Yearbook*, **3**, 117–149.

Stemberger, J.P. (1983). Speech errors and theoretical phonology: A review. *Technical Reports in Cognitive Science, Carnegie-Mellon University*.

Terbeek, D. (1977). A cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics*, **37**.

Lexical Neighborhoods in Speech Production: A First Report[1]

Stephen D. Goldinger and W. Van Summers[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This is a draft of a paper presented at the 117th meeting of the Acoustical Society of America, Syracuse, NY., May, 1989. This research supported by NIH Research Grant NS-12179-11.

[2] W. Van Summers is now with the Army Audiology and Speech Center, Walter Reed Army Center, Washington, DC. 20307.

Abstract

Investigations of speech production have shown that talkers will systematically alter the acoustic-phonetic properties of their utterances in response to changes in the context in which the words are spoken. Well-known examples of such contexts are the presence of a loud background noise, coginitive workload, or the linguistic context surrounding the target word in a sentence. Recent work by Balota, Boland, and Shields (1989) suggests that factors intrinsic to words, such as their frequencies, may also affect the durations of spoken words. The present paper reports the results of a preliminary investigation of the effects of similarity neighborhood structure on speech production. Global, as well as segmental, comparisons of subjects' productions of words from dense and sparse lexical neighborhoods will be presented.

Lexical Neighborhoods in Speech Production: A First Report

Research on both the production and perception of speech has long been represented by a search for the acoustic and phonetic invariants in the signal. By now, we may safely conclude that speech production is best described as a dynamic process, or an "open system". Our use of the term "open system" is intended to convey the idea that the characteristics of the physical signal in speech communication may be simultaneously modulated by numerous factors both intrinsic and extrinsic to the talker.

It has been known for many years that conditions *extrinsic* to the talker can affect speech production. A well-known example is the *Lombard effect*, the phenomenon originally reported by Lombard in 1911 (cited in Lane & Tranel, 1971), in which talkers increase their vocal effort in the presence of a loud background noise (see also Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). Beyond the findings that extrinsic ambient or physical conditions affect speech production, further research has shown that talkers modify their speech in accordance with the unique needs of the listener. For example, Moslin and Keating (1977) have shown that mothers tend to produce the stop consonants /b/ and /p/ with a greater voicing distinction when they speak to their young children than when they speak to adults or to older children.

In addition to the findings that conditions extrinsic to talkers affect speech production, several findings indicate that conditions *intrinsic* to talkers affect speech production as well. More specifically, there is evidence to suggest that variations in talkers' *representations* of language and linguistic communication may underlie changes in speech. For example, whereas Moslin and Keating (1977) noted that mothers alter their speech for the benefit of young ears, other research has shown that the intrinsic "information value" of a message affects the spoken quality of the utterance, without special reference to any particular ambient situation or kind of listener.

The most notable example of such an effect was reported by Lieberman (1963), who found that words spoken in highly predictable contexts are less intelligible when removed from context than words spoken either in isolation or in less predictive sentence contexts. Furthermore, Lieberman found that words were more precisely articulated upon first production than upon subsequent productions (the distinction that has come to be known as the *given* vs. *new* distinction). This experimental manipulation reflects Lieberman's working assumption that words lose some of their communicative value when they are spoken in highly redundant contexts, such as in cliches, or when one simply repeats the same word over and over. Following this reduction of information inherent to the word in a particular linguistic context, articulation of the word becomes less precise. More generally, Lieberman has shown that talkers decrease articulatory effort in production when the linguistic or semantic context of the utterance itself provides enough information to keep the words' meaning robust to phonetic inconsistencies. Similar experimentation on this *redundancy effect* has been performed by Charles-Luce (1987), who studied neutralization in both given and new words and found results that corroborate Lieberman's earlier findings.

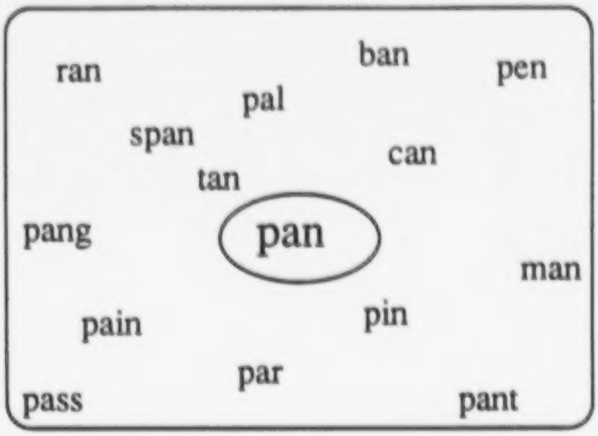
Beyond the speaker's semantic or pragmatic interpretations of the linguistic context, recent evidence suggests that more subtle linguistic variations can also affect speech production. Recent naming studies by Balota and his colleagues (Balota & Chumbley, 1985; Balota & Shields, 1988; Balota, Boland, & Shields, 1989) suggest that a talker's *familiarity* with different words may have direct effects on their production. The common finding in the word recognition literature is that subjects in a naming task respond to high frequency words more rapidly than low frequency words. Balota and his associates have replicated the classic word frequency findings in their recent work, but have also included additional acoustic measures on their subjects' spoken responses. Balota et al. (1989) observed not only reliable reaction time differences for subjects to *initiate* responses to high and low frequency words, but they observed reliable differences in spoken word *durations* as well. Not only was phonation initiated later for low frequency words, but, once initiated, the durations of low frequency words were longer than the durations of high frequency words. Balota et al. suggested that this result may reflect different degrees of familiarity with the articulatory motor programs necessary to produce rare and common words.

In the present study, we were interested in examining the effects of another intrinsic characteristic of words on their production. Specifically, we investigated the effects of words' *similarity neighborhood* structures on production. A similarity neighborhood is defined simply as a group of words that sound similar to any given word (see, e.g., Luce, 1986). The basic property of similarity neighborhoods that we examined in the present study is *neighborhood density*, which refers to the total number of words resident in the referent word's neighborhood. An example of two neighborhoods with differing densities is shown in Figure 1:

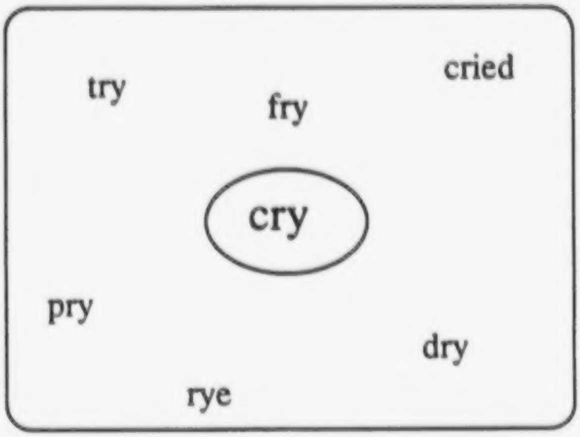
Insert Figure 1 about here

As these schematic neighborhoods show, some words have many neighbors while others have few neighbors. For convenience, we refer to these kinds of neighborhoods as "dense" and "sparse," respectively.

It has been observed in several spoken word recognition experiments that neighborhood density affects the speed and accuracy of recognition (Luce, 1986; Goldinger, Luce, & Pisoni, 1989). In experiments using a variety of tasks, it has been shown that words from sparse neighborhoods are recognized more quickly and accurately than words from dense neighborhoods. One might expect that neighborhood density may affect speech *production* as well, and the reasons for such an expectation are closely related to the perceptual consequences of large neighborhoods. For instance, in dense neighborhoods, small variations in articulation


ran
ban
pen
pal
span
can
tan
pang
pan
man
pain
pin
par
pass
pant

Dense Neighborhood


try
cried
fry
cry
pry
dry
rye

Sparse Neighborhood

Figure 1. Schematic illustration of dense and sparse similarity neighborhoods.

may result in the phonetic realization of an unintended word. In general, this is less likely to occur in sparse neighborhoods. Accordingly, it may be reasonable to hypothesize that the articulation of words from dense neighborhoods proceeds more slowly, carefully, or deliberately than the articulation of words from sparse neighborhoods. This prediction entails only the observation that dense neighborhoods provide a more narrow margin of error (for either word perception or production), and the assumption that the talker has some tacit representation of these limitations available during production. The present study attempts to address this assumption.

In order to assess the influence of lexical neighborhood density on speech production, we employed a *repetition* paradigm. Minimal pairs of words, differing only in voicing of initial consonants, were selected from both sparse and dense neighborhoods. Subjects were asked to read these pairs aloud and acoustic measurements were taken from their utterances. We were primarily examining the degree of contrast within pairs, with special attention to differences in VOT (voice onset time). We considered several predictions. First, we predicted that the VOT differences between words in pairs from dense neighborhoods would be larger than the VOT differences between words in pairs from sparse neighborhoods. Our second prediction was based on the Lieberman (1963) experiments, in which it was demonstrated that redundant contexts produce reductions in phonetic contrasts. Following Lieberman's findings, we hypothesized that if our subjects were required to produce minimal pairs repeatedly, some reduction of the phonetic contrast within the pairs would occur across trials. However, we did *not* expect this reduction to occur to equivalent degrees for pairs from both dense and sparse neighborhoods. Instead, we predicted that, across trials, the absolute VOT difference between the members of pairs from *sparse* neighborhoods would decrease considerably, whereas the difference between members of pairs from *dense* neighborhoods would decrease less, or perhaps remain unchanged. Our third, related, prediction was that the acoustic-phonetic characteristics of spoken words from dense neighborhoods would be *less variable* across repetitions than the characteristics of spoken words from sparse neighborhoods. This prediction was again derived from the assumption that the articulatory motor programs associated with words from dense neighborhoods may be more constrained than those associated with words from sparse neighborhoods.

Method

Stimuli. The stimuli for this experiment were 12 minimal pairs of words, distinguished within each pair by differences along the dimension of initial consonant voicing, such as the minimal pair *dutch - touch.* Four of the twelve pairs of words contrasted labial stops, four pairs contrasted alveolar stops, and four pairs contrasted velar stops. Half of these sets of minimal pairs were from dense neighborhoods, and half were from sparse neighborhoods. The order of words within each of these pairs was counterbalanced, yielding a total of 24 minimal pairs for presentation to subjects.

Procedure. The experiment was conducted in three blocks. Each block contained sixteen presentations of each of eight minimal pairs. In any given block of presentations, the phonetic contrast remained constant. That is, one block consisted entirely of /b/ - /p/ or /p/ - /b/ pairs, and so on. The pairs were presented in random order at a fixed rate of one pair every two seconds. The subject's task was to read each pair as it appeared on a CRT monitor, and to produce each pair in a natural manner. This procedure lasted approximately one hour and generated 384 tokens.

Results and Discussion

Two main findings were obtained. The first set of data pertain to word-initial VOT, and are shown in Figure 2:

Insert Figure 2 about here

Figure 2 shows the averaged absolute differences in VOT for all pairs of words. These data were computed by measuring the duration of VOT for the first word in any minimal pair and subtracting the duration of VOT for the second word of the pair from this initial value. This figure displays the averaged absolute values of these differences for pairs from dense and sparse neighborhoods. Because we were interested in the possibility that the magnitude of these differences might change over trials, the data are plotted in two halves. Recall that all pairs were spoken 16 times over the course of the experiment. The left side of the figure shows the mean differences observed across the first eight repetitions of all pairs. The right side shows the mean differences observed across the last eight repetitions.

Two effects that deserve mention are shown in Figure 2. First, there was a main effect of neighborhood density; the VOT differences within word pairs from dense neighborhoods was larger than the difference within word pairs from sparse neighborhoods. While this finding *was* statistically significant [$F(1,380) = 6.80, p < .01$], it should be noted that the difference was confined to only the pairs spoken in the second half of the session. This observation relates to the second interesting result. There was a significant interaction of neighborhood density and trials [$F(1,380) = 6.58, p < .02$]. As the talker repeated these minimal pairs over and over, the distinctions between words in the pairs from sparse neighborhoods became smaller, but the distinctions between words in the pairs from dense neighborhoods actually became slightly *larger*.

A second set of measurements performed on the utterances are shown in Figure 3:

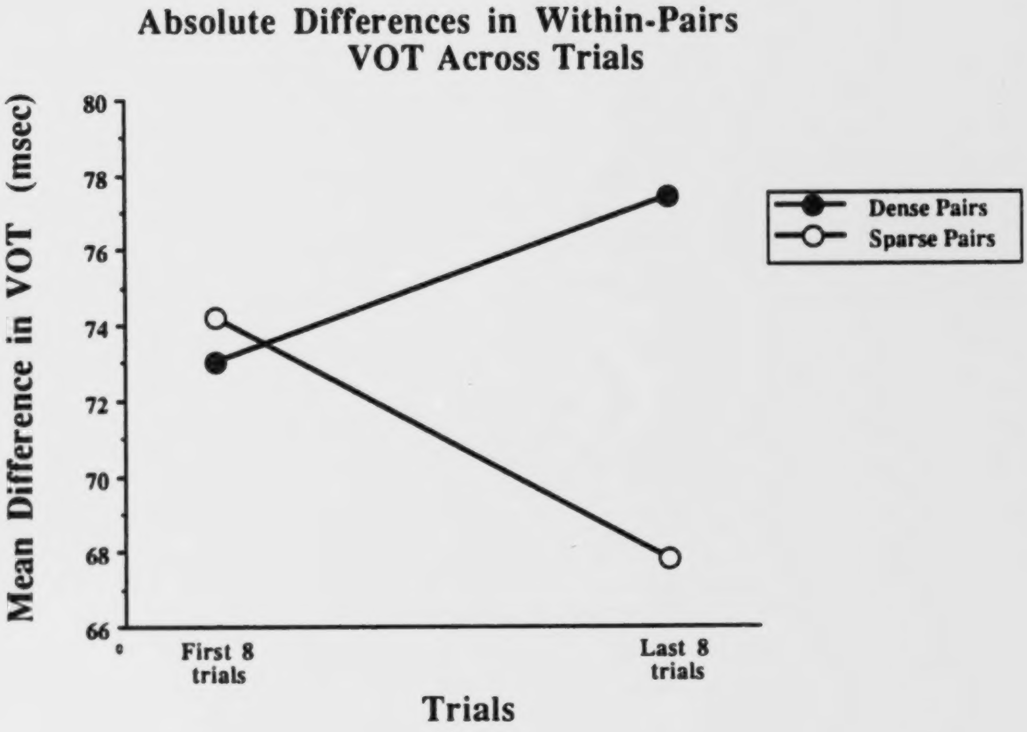


Figure 2. Absolute differences in VOT durations within minimal pairs of words as a function of neighborhood density and number of repetitions.

Insert Figure 3 about here

Figure 3 shows the mean durations of the *inter-word intervals* produced within pairs, across all trials. This value is simply the duration of the silent period that occurred between the offset of the first word and the onset of the second word within each minimal pair. As the figure shows, we obtained a large main effect of neighborhood density on the durations of these intervals $[F(1,380) = 71.65, p < .001]$. The pause durations between words in pairs from dense neighborhoods are, on the average, 53 ms longer than the pause durations between words in pairs from sparse neighborhoods. Although, as the figure shows, there was a decrease in the average magnitude of these interval durations across trials, the implied interaction was not statistically reliable.

Taken together, the present findings regarding VOT changes across repetitions and pause durations suggest that similarity neighborhood densities in the mental lexicon can affect speech production. One potential explanation for such effects relates to the particular motor programs associated with the words in memory. As discussed above, greater constraints may be required for the articulation of words from dense neighborhoods, simply because small variations in the acoustic-phonetics of such words are more likely to result in the production of an unintended word. Accordingly, one might expect minimal pairs from dense neighborhoods to display less sensitivity to the repetition manipulation employed in the present study.

Although we have only described articulatory constraints as a possible explanation for these findings, another potential account is available as well. In keeping with the spirit of the Lombard effect and the observations of "motherese" reported by Moslin and Keating (1977), one might speculate that words from dense neighborhoods are more carefully articulated only for the benefit of the *listener*. Essentially, this claim implies that the speaker "knows", in some tacit sense, which words are from dense neighborhoods, and also that words from dense neighborhoods are more difficult for listeners to recognize. The differing strictures of the production constraints applied to words from dense and sparse neighborhoods would then mirror the listener's needs. The same explanation could also be applied to the differences in word durations, noted by Balota et al., for high and low frequency words. Although a "communication-based" explanation such as this is *consistent* with the present findings, it rests upon several assumptions that may or may not be reasonable. Primarily, this sort of explanation requires that speakers have tacit awareness of all the frequencies and relative neighborhood densities of all the words in their lexicons, and that this information is continuously accessed throughout speaking. On the other hand, an articulatory/motor-based explanation requires a far simpler mechanism– the degrees of freedom associated with producing intended versus unintended words in different-sized neighborhoods, combined with the weak assumption that speakers try to produce intended words. Given the data available at present, the articulation-based explanation appears more parsimonious and reasonable.

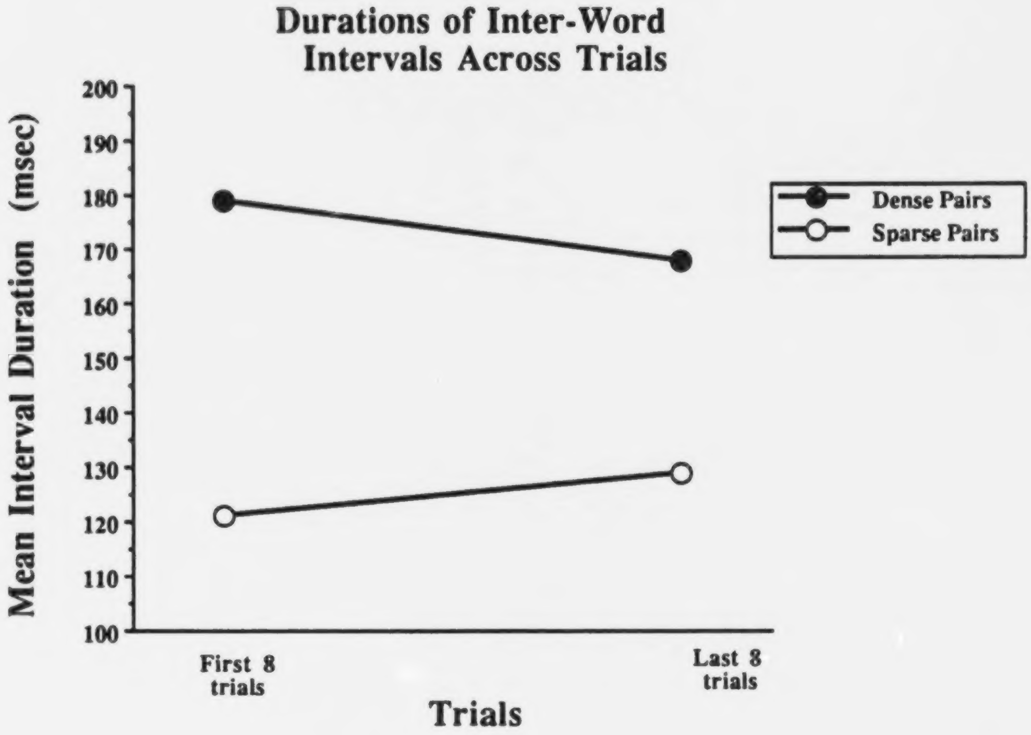


Figure 3. Absolute differences in inter-word intervals within minimal pairs as a function of neighborhood density and number of repetitions.

While the findings related to VOT contrasts and pause durations within pairs suggest that neighborhood densities can indeed affect word production, we should mention several analyses that did *not* reveal any reliable differences between words from sparse and dense neighborhoods. Some of the measures we have not detailed in this report include degrees of variability in durations of VOT and pauses, and degrees of variability in F1 and F0. We had considered the hypothesis that temporal or spectral variability would be greater for pairs from sparse neighborhoods than for pairs from dense neighborhoods, again because of the stricter production constraints that may apply for words from denser neighborhoods. These predictions were not supported by our measurements. Similarly, we considered the possibility that if the speaker were actually trying to accentuate the differences within pairs for the *listener's* benefit, there would be a tendency for the speaker to vary the intonation pattern more for pairs from dense neighborhoods, so the pairs would sound like *bill - PILL*. Again, we did not find this pattern in our data (which may be considered further evidence in favor of an articulation-based explanation of these results, as opposed to a communication-based theory).

In summary, the present data suggest that lexical neighborhood densities can affect speech production. Although the effects observed were small, there is enough consistency in the data to warrant further research on the role of lexical organization in speech production. Simply collecting more data from more speakers is certainly a first necessary step, but may not be sufficient in and of itself. Alternate measures may provide a cleaner, more comprehensive account of the phenomenon. The measurements made in the present study were focused primarily on VOT in word initial stops. Clearly, there are many other aspects of the speech signal that could have been examined as well. With more extensive study using a wider range of dependent measures, we hope to learn more about how the organization of the mental lexicon affects the production of lexical items.

References

Balota, D.A., Boland, J.E., & Shields, L.W. (1989). Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, **28**, 14-36.

Balota, D.A., & Chumbley, J.I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, **24**, 89-106.

Balota, D.A., & Shields, L. (1988). Localizing word-frequency effects in pronunciation. Paper presented at the twenty-ninth annual meeting of The Psychonomic Society, Chicago, Ill.

Charles-Luce, J. (1987). The effects of semantic context on voicing neutralization. *Research on speech perception progress report no. 13*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Lane, H.L., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, **14**, 677-709.

Lane, H.L., Tranel, B., & Sisson, C. (1970). Regulation of voice communication by sensory dynamics. *Journal of the Acoustical Society of America*, **47**, 618-624.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172-188.

Lombard, E. (1911). Le signe de l'elevation de la voix. (Cited by Lane & Tranel, 1971).

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Moslin, B.J., & Keating, P.A. (1977). Mothers' simplification of phonetic input to their children in English and Polish. *Journal of the Acoustical Society of America*, **61**, **supp. 1**, D7.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, **84**, 917-928.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

On the Perceptual Representation of Vowel Categories[1]

Keith Johnson

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported by NIH Training Grant No. NS-07134-11. I appreciate the comments of Bob Port, John Logan, Steve Goldinger, Bob Bernacki, Dawn Behne and Scott Lively. Denise Beike assisted in conducting the experiment. This paper was presented at the 118th meeting of the Acoustical Society of America, St. Louis, MO. (JASA 86:S100).

Abstract

This paper presents the results of a study in which vowel production and perception were compared in two ways. First, the perceptual vowel space (as indicated in a 'direct prototype estimation' experiment) was compared with the F1, F2 space of vowels produced by the same subjects. In this experiment, all of the subjects selected prototypes which have formant values between those of vowels produced by male and female speakers. The second type of comparison involved testing two types of vowel spectrum representation - whole-spectrum representation and spectral-peak representation. Whole spectral templates constructed from the subjects' vowel productions were better able to predict the perceptual data than could templates which contained information about the frequency and amplitude of spectral peaks.

On the Perceptual Representation of Vowel Categories

The preliminary study which is reported here was concerned with two important issues in vowel perception. The first issue is the relationship between the perceptual vowel space and the acoustic description of vowels. Most previous research concerning the perceptual vowel space has used multidimensional scaling (Fox, 1982; Pols, van der Kamp and Plomp, 1969; Singh and Woods, 1970; Terbeek, 1974), and it has been difficult to relate the perceptual dimensions to acoustic dimensions beyond statements to the effect that 'dimension 1 is correlated with F1 frequency', and so on. The method used here is an attempt to collect perception data which can be directly related to production. The second issue which I attempted to address has to do with the perceptual representation of vowel spectra. There has been some debate about the perceptual role of vowel formants or, perhaps more generally, spectral peaks (Bladon, 1982; Plomp, 1974). To address this issue, I compared two types of spectral representation - whole spectrum representations and spectral peak representations.

Three types of data were considered. First, vowel production data were collected. These data served as a reference point for the two other types of data. Perception data were collected in a task which I will call 'direct estimation of vowel prototypes' (see Samuels, 1982). These data can be described in terms which allow a direct comparison with vowel formant values and will, therefore, make it possible to study the relationship between the perceptual vowel space and the acoustic vowel space. The third type of data comes from a model study of vowel representation. In this study, two types of vowel representations were constructed from naturally produced vowels and then compared in terms of their relative success at predicting the perception data.

Production Data

Four subjects (2 male, 2 female) participated in the experiment. The subjects were college undergraduates who were linguistically unsophisticated, and speakers of the same dialect of American English. These four subjects participated in both the production and perception portions of the study.

In the production portion of the experiment, the subjects read 10 repetitions of the [hVd] words in a carrier phrase (see Peterson and Barney, 1952). Formant values during a steady-state portion of each vowel were measured using the autocorrelation LPC analysis with rootsolving for formants. Average formant values of the 9 monophthongal vowels are shown in Figure 1. Notice three features of this data: (1) vowels are unevenly distributed in the F1, F2 space, (2) vowels produced by the female subjects had generally higher formant values, and (3) separation between /a/ and /ɔ/ was maintained.

Insert Figure 1 about here

Perception Data

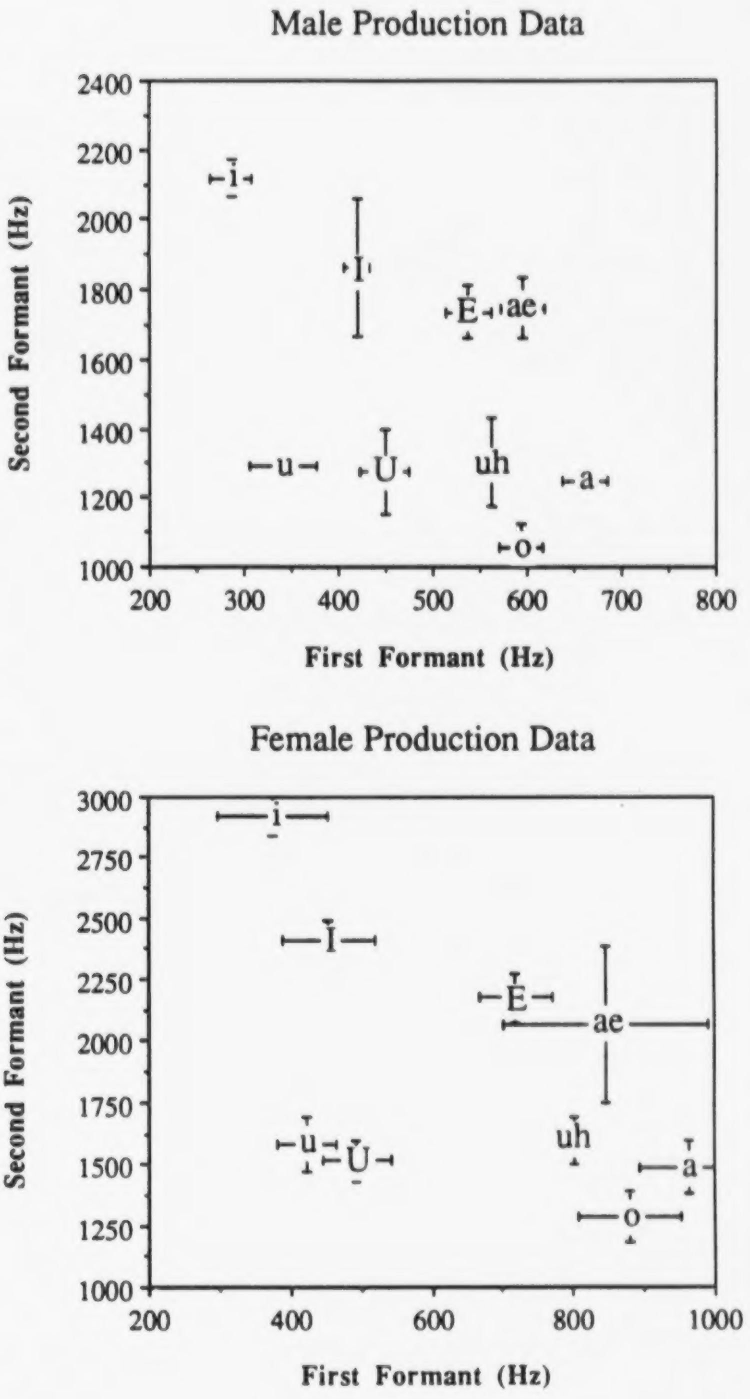
The perception portion of the study involved a task which was used by Samuels (1982). This task allows subjects to directly estimate the formant values of their vowel prototypes. Samuels' task was modified for the present experiment in two ways; first, subjects estimated two dimensions (F1 in F2), instead of one (Samuels' subjects estimated prototypical values of VOT), and second, step sizes for F1 and F2 were fixed. Samuels (1982) varied step size in order to insure that subjects would not rely on a step counting strategy in the task. In the current experiment, there were several differences which suggested that fixed step sizes could be used. Subjects estimated prototypes of 11 different phonemes (rather than two in Samuels' experiment) along two dimensions (instead of one). As will be noted below, the decision to use a fixed step-size complicated the interpretation of the results. The F1 and F2 values of the stimuli used in the task are shown in Figure 2. This is a plot of the F1 and F2 values of each token in an array of 305 synthetic syllables. The range of these formant values covers the ranges for both male and female speakers, and the tokens are evenly spaced in Bark. F3 was computed by a formula published by Nearey (1989). The tokens were 150 ms long with five steady-state formants.

Insert Figure 2 about here

In each trial, the subject saw one of the [hVd] words (in normal orthography) on a video monitor and heard the token indicated by a triangle in the figure. The subject's task was to adjust F1 and F2 until the synthetic token sounded like the vowel in the visually presented word. They adjusted F1 and F2 by pushing buttons labelled 'up', 'down', 'left' and 'right' to hear a new token from the array. F2 was increased by pushing the 'up' button and decreased by pushing the 'down' button. F1 was increased by pushing the 'right' button and decreased by pushing the 'left' button. Subjects had to rely on auditory information alone, as there was no visual indication of their location in the F1, F2 space.

Since the subjects were linguistically naive, they, at first, had no idea what to do, and a single trial could last as long as ten minutes. They had to explore the space and find the best examples of each vowel category. They soon learned, though, which buttons to press to find the appropriate area of the F1, F2 space for each vowel.

There were three conditions in the experiment, corresponding to three stimulus sets - low F0 (120 Hz), high F0 (240 Hz), and noise excited. Conditions were presented in same



Male Production Data

Female Production Data

Figure 1. Vowel formant measurements for the male and female subjects.

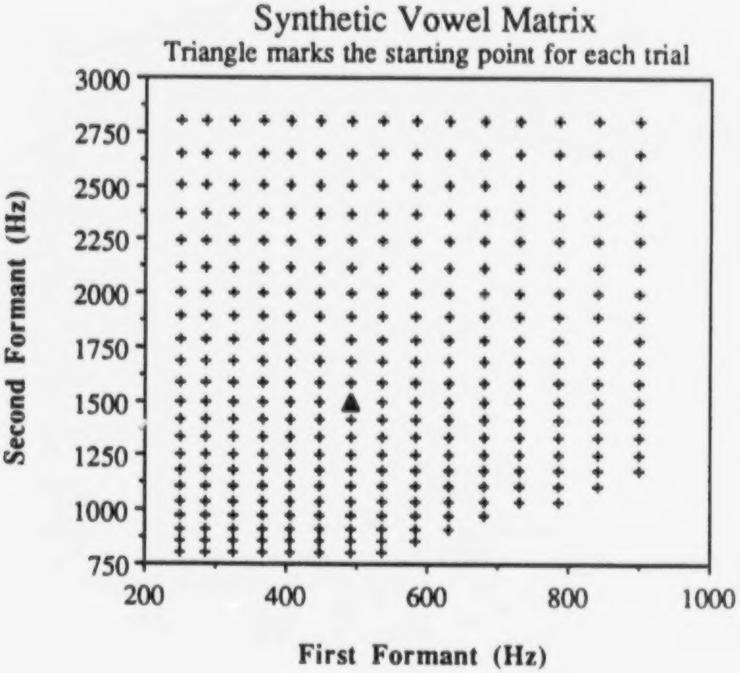


Figure 2. F1 and F2 of the synthetic tokens used in the perception experiment.

order (low, high, then noise) to each subject. Each of the four subjects estimated the vowel prototypes of the 11 vowels 10 times in each condition for a total of 330 observations per subject. The same four subjects who participated in the production portion of the study participated in this perception experiment.

Results averaged across subjects and conditions are shown in Figure 3. Notice in this figure: (1) vowels are more evenly distributed in the F1, F2 space. This seems to validate Liljencrants and Lindblom's (1972) suggestion that, all things being equal, hearers will prefer a vowel system in which the vowels are evenly distributed in the vowel space. The fact that this pattern was not found in production indicates that the constraint can be relaxed when other, nonspectral, cues are available. (2) /a/ and /ɔ/ are merged. This seems to reflect the confusion in this dialect of /a/ and /ɔ/ and parallels other cases in which production and perception are different for contrasts which are in the process of change).

Insert Figure 3 about here

The pattern seen in this figure was found in the data of all four subjects in all three stimulus set conditions. Figure 4 shows the data broken down by subject gender in the top panel and by condition in the bottom panel. The top panel indicates that the female subjects chose values which were slightly more extreme than those chosen by the male subjects. Note that they did not choose items with generally higher formant values. Thus, this data seems to be indicative of the subjects' diligence in the task more than anything else. The remarkable feature of the bottom panel (data broken down by condition) is the lack of an effect for F0 differences. At the present it is not clear whether this is an artifact. If subjects adopted a step-counting strategy, we would expect no difference between conditions. Therefore, in the next experiment in this line of research, the starting point for each trial will be varied.

Insert Figure 4 about here

Figure 5 shows the production and perception data plotted together. The estimated perceptual prototypes are plotted with solid symbols and the formant values for male and female productions are plotted with open symbols. Note that the perceptual prototypes (as measured in this task) have formant values between those of male and female productions of the same vowels. This raises the tantalizing possibility (subject to the methodological concerns mentioned earlier) that the perceptual categories for vowels are some sort of compromise between typical male and female vowels.

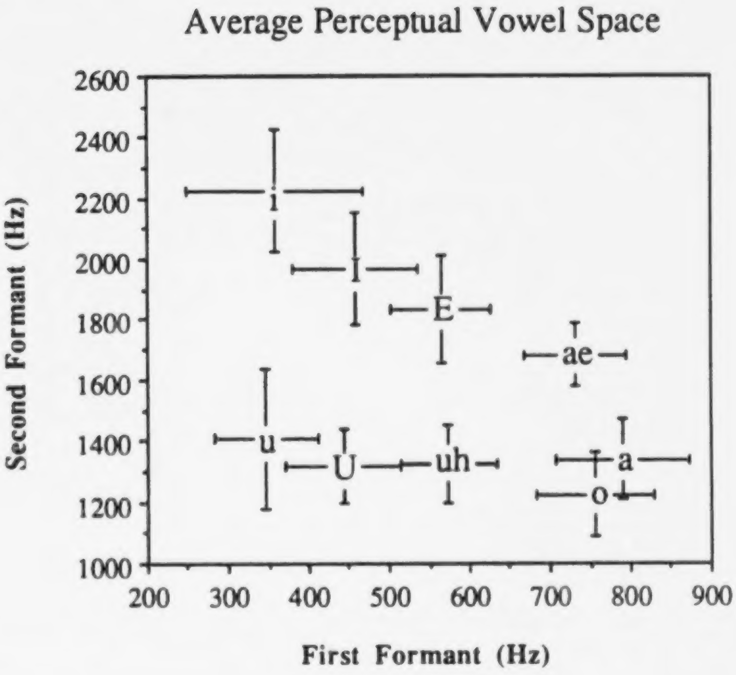
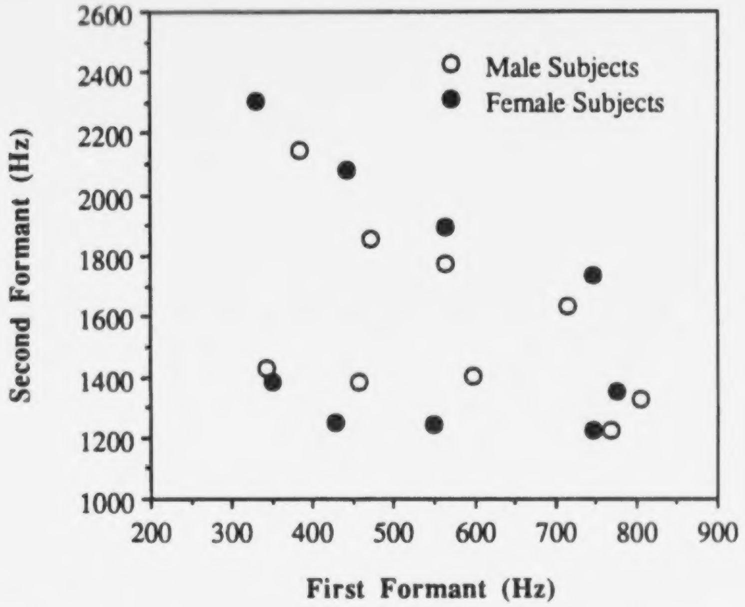


Figure 3. Prototype estimates averaged across subjects and conditions.

Perception Data - Broken down by subject gender



Perception Data - Broken down by stimulus set

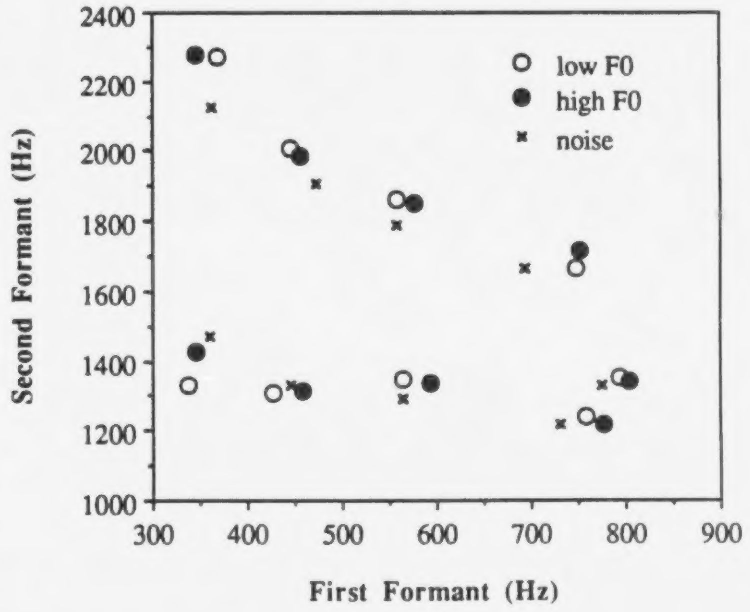


Figure 4. Prototype estimates by subject gender and condition.

Insert Figure 5 about here

Modeling Study

We now turn to the question of vowel representation. To investigate the perceptual representation of vowels, I attempted to model subjects' performance in the perception test by the use of pseudo-auditory vowel spectra generated from the subjects' vowel productions.

Figure 6 shows the steps involved in generating the auditory spectra. After calculating an FFT, the spectrum was filtered using the auditory filter shapes and bandwidths published by Patterson (1976). A smearing function was then applied to the spectrum. The smearing function was a Gaussian filter with a bandwidth of 3 Bark. This step in the construction of spectra is an attempt to model Chistovich, Sheikin and Lublinskaja's (1979) suggestion that spectral components are integrated over a 3 Bark range in speech perception. Finally, an equal loudness contour was applied to the spectrum (see Bladon and Lindblom, 1981).

Insert Figure 6 about here

Pseudo-auditory spectra of the subjects' productions were generated by analysing the same window of samples from which the formants were measured. Templates for the vowel categories were constructed by averaging across subjects (speaker dependent templates were also constructed and tested with generally poorer results than these cross-speaker averages). Auditory spectra of the synthetic stimuli which had been used in the perception test were also generated.

Figure 7 is an outline of the test of spectral representation. For each of the templates constructed from the subjects' productions, one of the synthetic stimuli was selected as the best match. Two types of representation were used to select the best match between production template and synthetic token. In one case, the Euclidean distance between spectra was calculated. This will be called the whole-spectrum model because each point in the spectrum contributed an equal weight to the distance measure. In the other case, spectral peaks were compared. Spectral similarity in this spectral-peak model was dependent upon the frequency and amplitude of peaks in the auditory spectra.

Insert Figure 7 about here

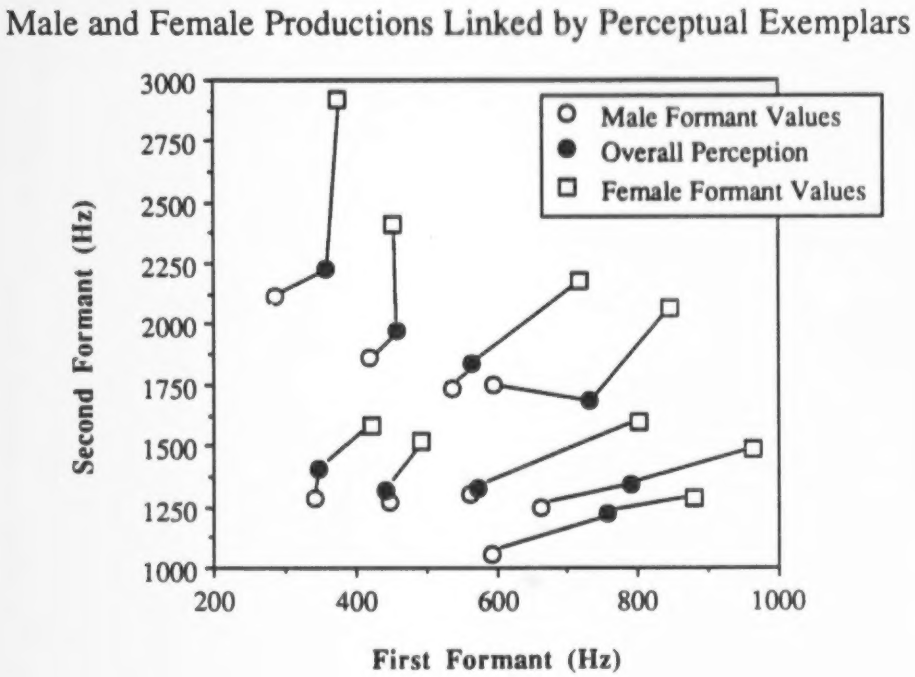


Figure 5. Average prototype estimates compared with vowel formant values of male and female speakers.

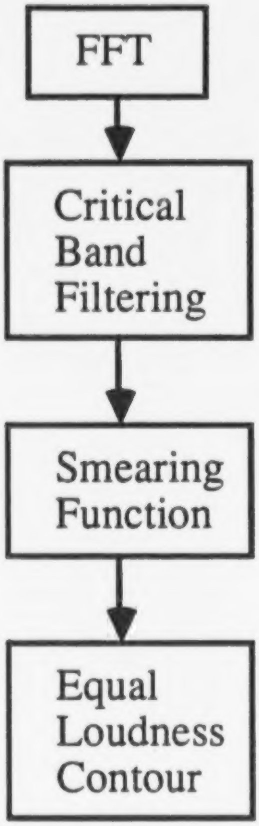

FFT
Critical Band Filtering
Smearing Function
Equal Loudness Contour

Figure 6. Outline of steps involved in producing the pseudo-auditory spectra.

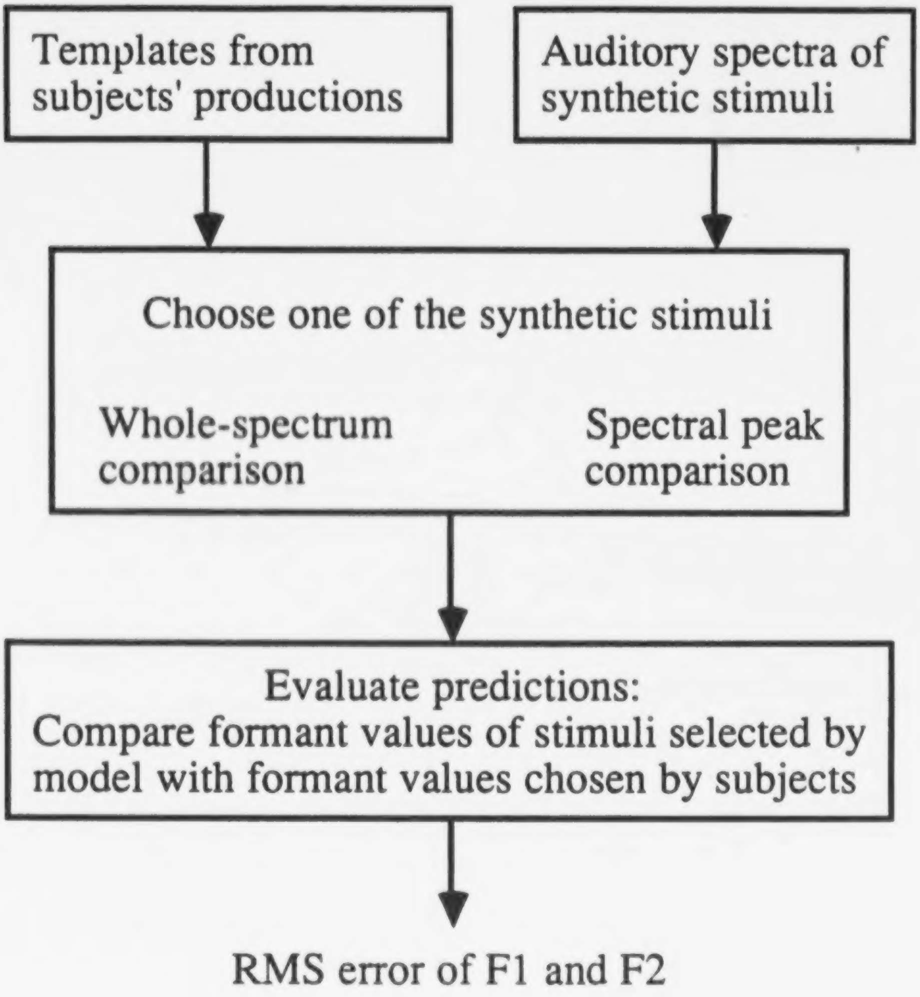

Templates from subjects' productions
Auditory spectra of synthetic stimuli
Choose one of the synthetic stimuli
Whole-spectrum comparison
Spectral peak comparison
Evaluate predictions:
Compare formant values of stimuli selected by model with formant values chosen by subjects
RMS error of F1 and F2

Figure 7. Flow chart of the procedure used to evaluate the two models.

The representations were evaluated by comparing the formant values of the tokens selected by template matching with the formant values of the tokens chosen by the subjects in the prototype estimation experiment. Root mean squared (RMS) error values (comparing the F1 and F2 of the tokens selected by the models with the average F1 and F2 selected by the subjects in the listening task) were calculated. Results of the study are shown in Figure 8. Plotted here are RMS error values for F1 and F2 (dark versus light bars) as a function of stimulus set (low F0, high F0, or noise excitation) and type of spectral representation (whole-spectrum, versus spectral-peak). The figure shows that the whole-spectrum matching produced predictions which were closer to the perceptual data than did spectral-peak matching, although it is also evident in this figure that neither method produced a remarkably close fit. Also, this figure indicates that the whole-spectrum approach produced predictions which were consistent across the different stumulus sets.

Insert Figure 8 about here

Conclusions

To summarize the present findings:

(1) The direct prototype estimation method produced data which reflect dialect characteristics and assumptions concerning the utilization of spectral space and so seems to provide useful data concerning the perceptual vowel space.

(2) There was little difference between the prototypes chosen for vowels with different F0 values (although it is possible that this is an artifact).

(3) The perceptual prototypes found in this experiment had formant values which were between the formant values of vowels produced by men and women.

(4) Whole-spectrum representations were better at predicting the perceptual data than spectral-peak representations.

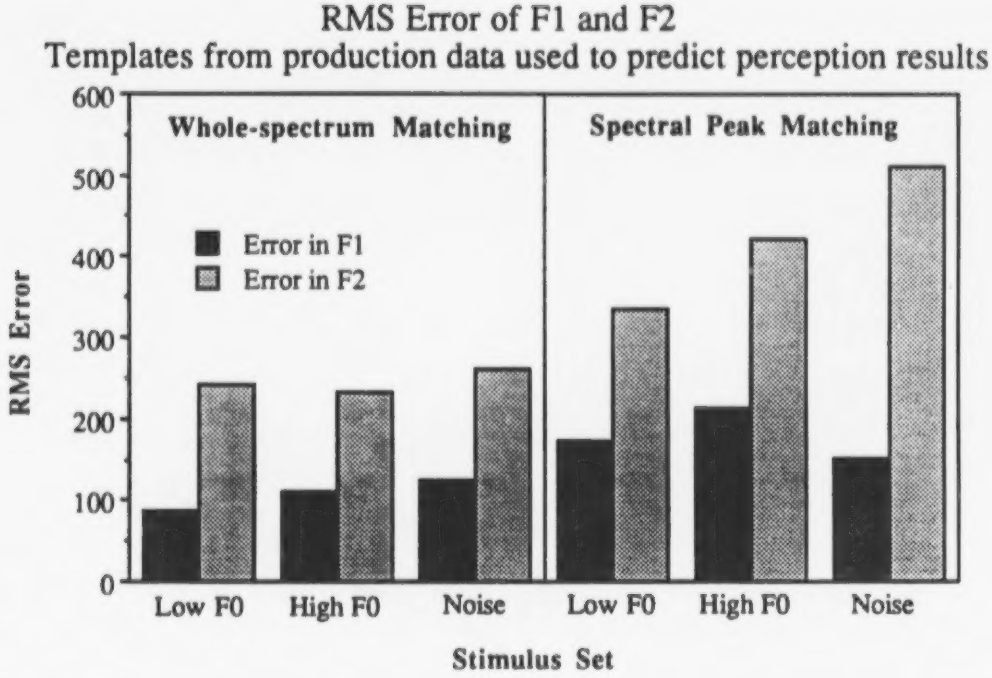


Figure 8. Results of the modelling study.

References

Bladon, R. (1982). Arguments against formants in the auditory representation of speech. In Carlson, R. & Granström, B., editors, *The representation of speech in the peripheral auditory system*. Elsevier Biomedical, Amsterdam.

Bladon, R. & Lindblom, B. (1981). Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America*, **69**, 1414–1422.

Chistovich, L., Sheikin, R., & Lublinskaja, V. (1979). 'Centres of Gravity' and spectral peaks as the determinants of vowel quality. In Lindblom, B. & Öhman, S., editors, *Frontiers of speech communication research*. Academic Press, London.

Fox, R. A. (1982). Individual variation in the perception of vowels: Implications for a perception-production link. *Phonetica*, **39**, 1–22.

Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, **48**, 839–862.

Nearey, T. (1989). Static, dynamic and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088–2113.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, **59**, 640–654.

Plomp, R. (1976). *Aspects of tone sensation: A psychophysical study*. Academic Press, London.

Pols, L., van der Kamp, L. & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America*, **46**, 458–467.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307–314.

Singh, S. & Woods, G. (1970). Perceptual structure of 12 American vowels. *Journal of the Acoustical Society of America*, **49**, 1861–1866.

Terbeek, D. (1977). A cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics*, **37**, 1–271.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Glottal Effects on LPC Estimation of F1[1]

Keith Johnson

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

---

[1] I appreciate comments on an earlier version of this manuscript offered by Van Summers, David Pisoni and Bob Bernacki. The research reported here was supported by NIH Training Grant NS-07134-10 to Indiana University.

Abstract

Linear Predictive Coding (LPC) analyses of synthetic speech tokens are reported. Results indicate that both F0 and glottal spectral tilt affect the LPC estimate of F1. The practical implications of these effects for acoustic-phonetic research are discussed.

Glottal Effects on LPC Estimation of F1

This paper reports the results of a set of Linear Predictive Coding (LPC) analyses of synthetic vowels which were carried out in an attempt to determine the extent and nature of glottal influence on LPC estimates of F1. Fitch's (1989) recent criticism of the use of LPC by Summers et al. (1988) has cast doubt on the utility of LPC analysis in acoustic phonetics. Fitch (1989) was primarily concerned with the influence of glottal factors on LPC estimates of F1. Since LPC analysis is used extensively in the study of speech acoustics it is important to understand the effects of source characteristics (such as F0 and glottal tilt) on LPC estimates of F1.

The Synthesizer

A version of the Klatt (1980) cascade/parallel formant synthesizer was used. The synthesis program was written in the C programming language by Dennis Klatt (see Klatt and Klatt, 1990). This version of the program includes several changes in the voice source which make it possible to control spectral tilt and the proportion of the voice period during which the glottis is open (the "open quotient"). These changes will be briefly described here.

The synthesis program has two voice sources. One is the impulse source which was described in 1980. The other is a more natural source which was first proposed by Rosenberg (1971, see Klatt and Klatt, 1990, p. 838). One of the chief differences between the two voices is that the natural voice source has a definite closing time, and thus, more high-frequency energy. Open quotient is manipulated explicitly in the natural source and implicitly in the filtered-impulse source. The nominal control of open quotient is achieved in the impulse source by use of a critically-damped second-order filter with frequency equal to 0 and bandwidth proportional to the open quotient.[1] The effect of changes in the value of the open quotient is limited to frequencies below 500 Hz as illustrated in Figure 1(a). In addition to the open quotient, the newer version of the synthesizer includes a parameter for the manipulation of the tilt of the voicing spectrum (of either voice). A soft one-pole low-pass filter reduces energy in the higher frequencies while leaving energy below 500 Hz relatively unchanged. Figure 1(b) illustrates the spectral effect of this parameter.

Insert Figure 1 about here

---

[1] The time constant of the filter is determined by its bandwidth, so the temporal decay of the impluse source is manipulated nominally by the bandwidth of the glottal filter.

Figure 1(a) The effect of Open Quotient.

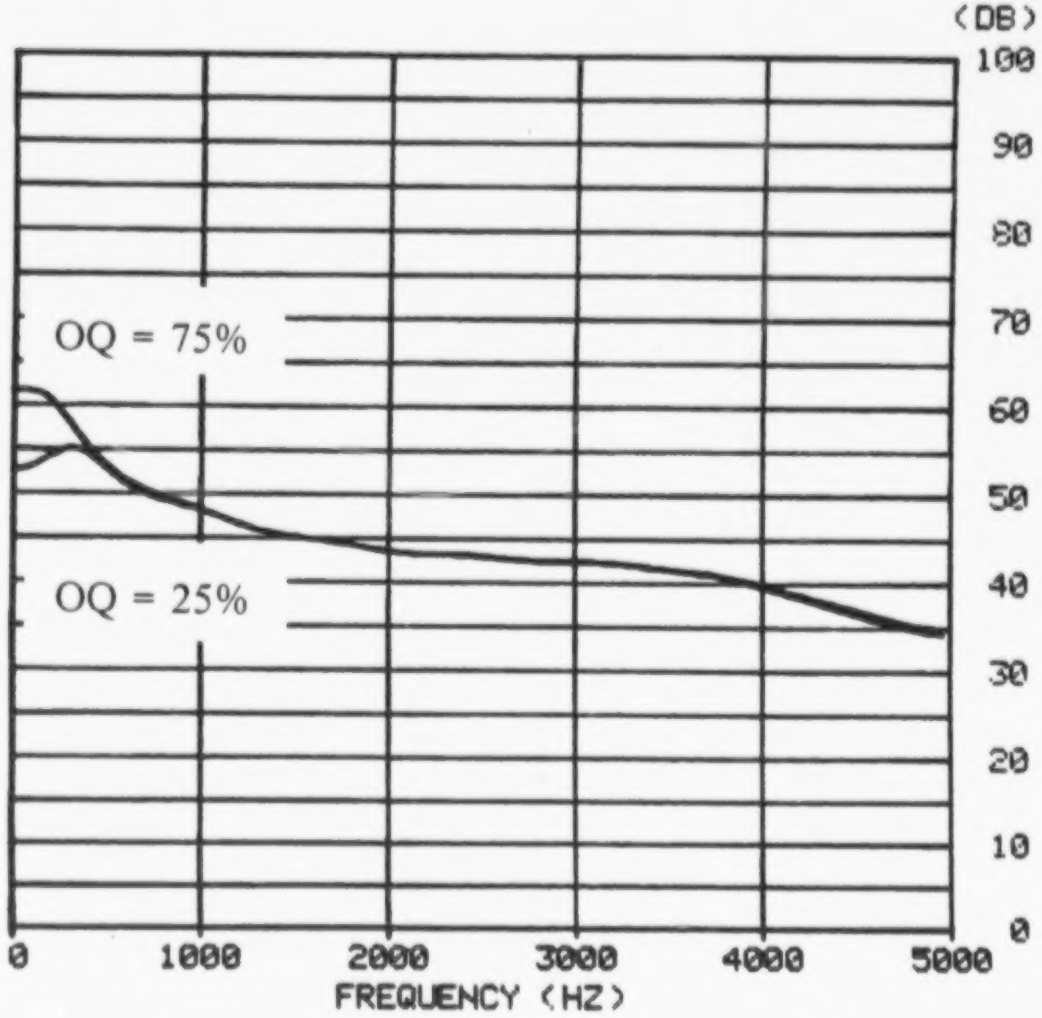


Figure 1. (a) The effect of changing the synthesizer open quotient. (b) The effect of changing the synthesizer tilt parameter. (c) The effect of changing preemphasis during LPC analysis

Figure 1(b) The effect of Glottal Tilt

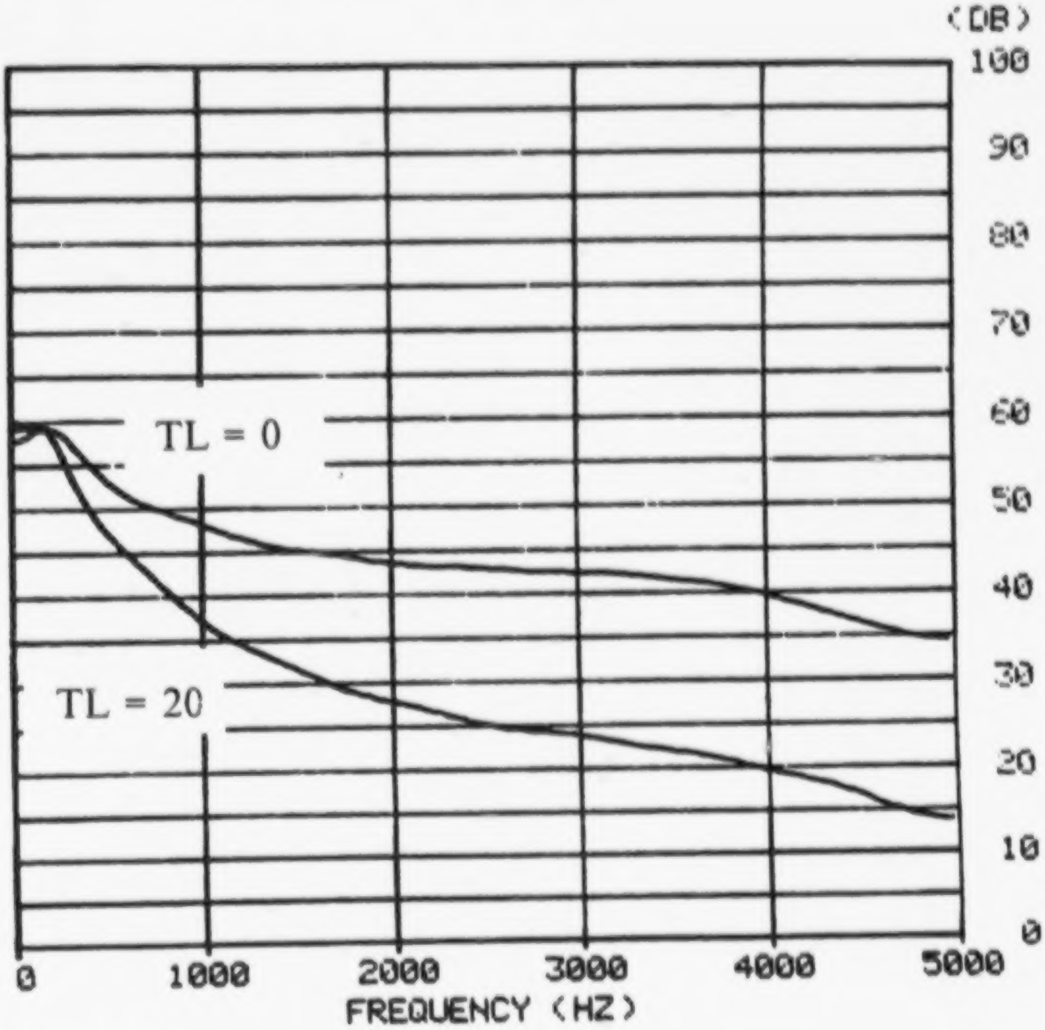
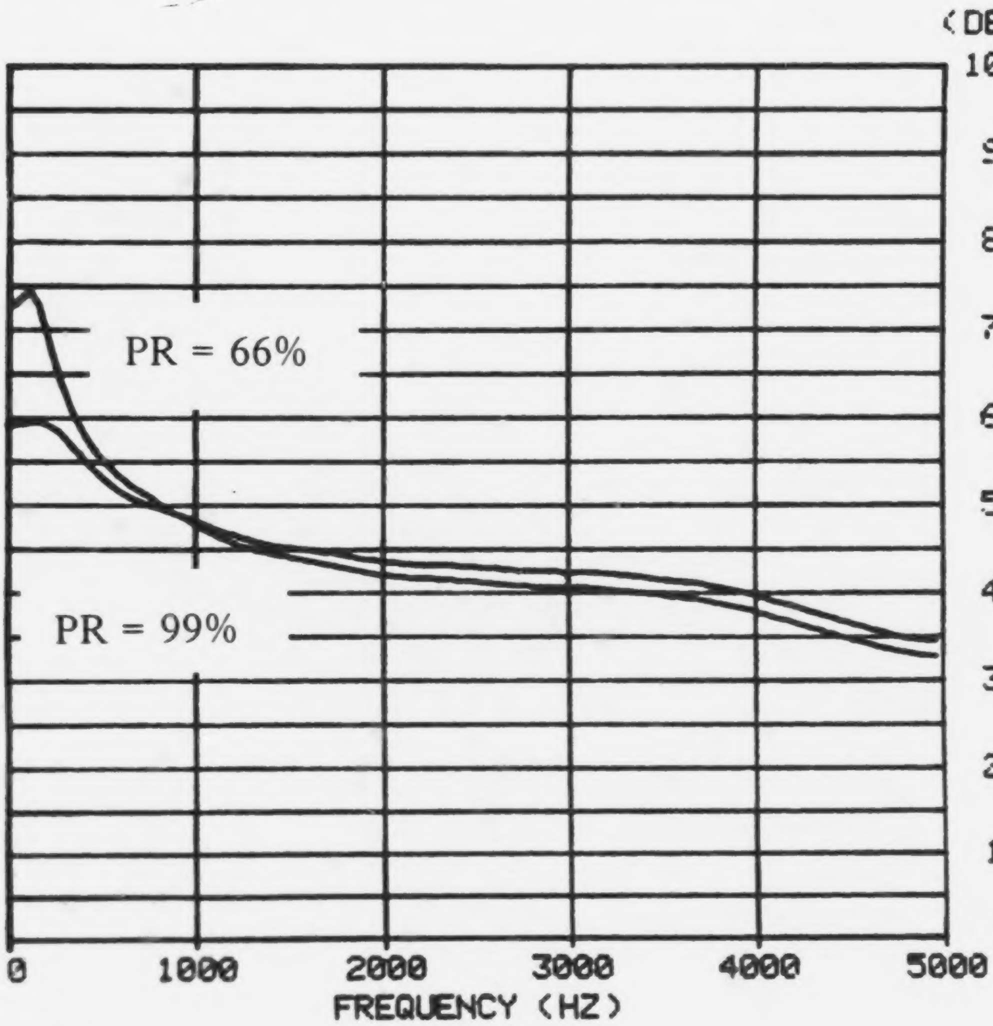


Figure 1(c) The effect of Preemphasis



F0 effects on LPC estimation of F1

Atal and Schroeder (1974) found that LPC estimates of F1 fluctuate around the actual F1 as a function of F0, and that high values of F0 showed greater fluctuation. This effect was also found in the present study. In addition, the data reported here suggest that the LPC estimate of F1 is drawn toward the strongest harmonic of the fundamental in the F1 region.

In the present study, F1 was estimated for a set of tokens which were synthesized with constant formant values (F1 500 Hz, F2 1500 Hz, F3 2500 Hz) and F0 ranging from 100 to 400 Hz in 10 Hz steps. The filtered-impulse source (with an open quotient of 50% and no added tilt) was used. The sampling rate of the tokens was 10 kHz. LPC coefficients (14) were calculated by the autocorrelation method and formants were estimated by peak-picking. Preemphasis was 99% and a 256 point Hamming window was applied to the waveform before the coefficients were estimated.

Insert Figure 2 about here

As indicated in Figure 2, F0 did have an effect on estimated F1. The estimate of F1 is drawn toward the strongest harmonic in the F1 region. Atal and Schroeder (1974) suggested that the decreased accuracy of LPC estimates of formant values (as F0 is increased) is a result of periodicity in the LPC residual signal. "The prediction error is periodic and exhibits strong correlation at delays equal to a pitch period and its multiples. Such correlations can introduce errors in the predictor coefficients" (p. 29). Because the pitch period for high F0 is smaller, the relative proportion of correlated speech samples is greater and thus, the amount of error in the predictor coefficients is greater for high F0 than for low F0. Atal and Schroeder suggest two methods to overcome this problem. First, they suggest that, "the errors due to voice periodicity can be completely avoided if the predictor memory is made large enough to include at least one period of the signal" (p. 29). Since the pitch period for even very high F0 is a good deal larger than the normal number of coefficients used in LPC analysis (for instance the pitch period for F0 of 400 Hz is 25 samples), Atal and Schroeder recommended the use of two sets of predictors - one for the region from 0 to 1000 Hz, and the other for the frequency region above 1000 Hz. Predictors of the first frequency region would be calculated from a down-sampled series of data points to allow for the desired increase in predictor memory.

Their second suggestion involves the fact that "prediction error is generally very large at the beginning of every pitch period." They suggest that "the interval over which the prediction error is minimized" be limited to "portions of the pitch period where the prediction error is relatively small" (p. 30). This suggestion was implemented quite directly in the

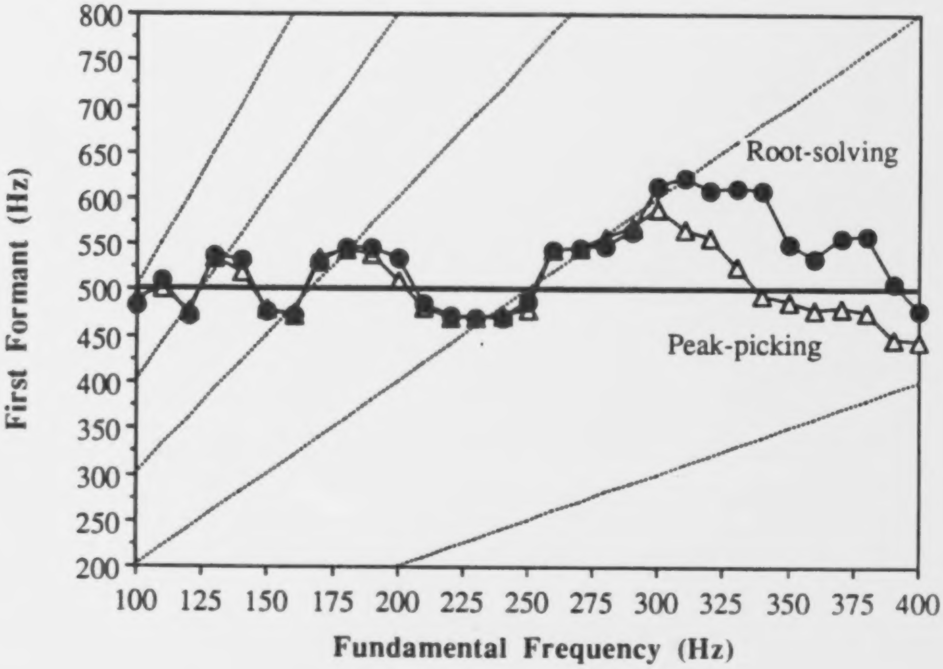


Figure 2. LPC estimates of F1 for synthetic stimuli with F1 of 500 Hz and a range of F0's from 100 to 400 Hz. Harmonics are indicated with dashed lines. The filtered-impulse voice-source was used (open quotient of 50%, no added tilt).

"sample selective LPC" method of Miyoshi et al. (1987), and more indirectly in the various techniques of pitch-synchronous LPC analysis (of which the most promising for phonetic research is the two-channel analysis technique of Krishnamurthy and Childers (1986)).

The finding reported here is that the LPC estimate of F1 is drawn toward the strongest harmonic in the F1 region. However, it should be noted that listeners' perceptual estimates of F1 display a similar pattern (Darwin and Gardner, 1985 and Assman and Nearey, 1987). In fact, Assman and Nearey report that of all the techniques that they used to predict listeners' estimates of F1 from the vowel spectrum, peak-picking from the power spectrum of the LPC coefficients corresponded most closely with the perceptual data. So, although it can be demonstrated that LPC estimates of F1 decrease in accuracy as F0 increases (which is quite definitely undesirable for the acoustic phonetician), it should be borne in mind that human perception seems to run up against the same limits in resolution that are faced by LPC analysis (see also Ryalls and Lieberman, 1982).

Spectral tilt effects on LPC estimates of F1

Makhoul and Wolf (1972) reported that the spectral tilt change produced by preemphasis (taking the first difference of the signal) results in an increase in LPC estimates of F1. Markel and Gray (1976) claimed that the tilt change produced by preemphasis (as normally used in LPC analysis) provides a more accurate representation of the vocal tract filter function (than no preemphasis at all) by compensating for the -6dB/oct tilt of the voice source. Since glottal tilt varies in speech production, it is important to know how naturally occurring variation in spectral tilt affects LPC estimates of formant values.

Summers et al. (1989) argued that the spectral tilt change produced by preemphasis is qualitatively different from the spectral tilt change produced by a change in glottal source function. This point is illustrated in Figure 1(b) and 1(c). Preemphasis affects (almost exclusively) the frequency region below 1000 Hz, while (un)rounding the corners of the glottal function affects higher frequencies.[2] It also should be noted that the spectra shown in Figure 1(c) represent a small change in preemphasis (from 99% to 66%). The difference between full preemphasis (differencing) and no preemphasis is over 20 dB in the region of F1. The spectral effect of preemphasis has been noted before (see Makhoul and Wolf, 1972 and Wong, Hsiao and Markel, 1980), but a comparison of preemphasis with natural tilt changes has not. The simulation reported below was carried out as a means of comparing the effect on LPC formant tracking of tilt changes produced by preemphasis with those produced by "rounding the corners" of the glottal source function.

---

[2] Changing the open quotient (which is normally correlated with (un)rounding the corners of the glottal function) affects low frequencies (below 500 Hz) almost exclusively, so in naturally produced tilt changes we expect to see change across the entire spectrum – as reported by Summers et al. (1989).

The effect of these two sources of spectral tilt on LPC estimates of F1 was investigated by synthesizing a set of neutral vowels (F0=100, F1=500, F2=1500, F3=2500 Hz) with four different levels of tilt (0, 10, 20 and 30 dB drop over the first 3000 Hz). One set of four tokens was synthesized using the filtered-impulse source and one set of four tokens using the natural voice source. LPC coefficients were then calculated with four different levels of preemphasis (99, 66, 33 and 0%).[3] The average LPC estimates of F1 are plotted in Figure 3 as a function of the measured tilt of the LPC spectrum.[4] Note that this combination of F0 and F1 (in Figure 2) resulted in an LPC estimate of F1 which was some 20 Hz below the actual F1. This magnitude of error is reflected in the F1 estimates shown in Figure 3. The LPC estimates of F1 were entered into an analysis of variance with factors; voice (impulse or natural), tilt (0, 10, 20 or 30) and preemphasis (99, 66, 33, 0%). Both the preemphasis and tilt factors produced a reliable effect on estimated F1 $[F(3,9) = 21.15, p < 0.01]$ and $[F(3,9) = 28.36, p < 0.01]$, respectively. As spectral tilt increased the LPC estimate of F1 was pulled down.

Insert Figure 3 about here

Interestingly, a change in glottal tilt had a different effect on the estimate of F1 than did a change in tilt produced by changing preemphasis (the interaction between preemphasis and glottal tilt was significant $[F(9,9) = 6.78, p < 0.01]$). The same degree of spectral tilt change produced by a change in glottal function produced a smaller change in estimated F1 (for moderately tilted spectrum; -4 to -7 dB/oct) than it did when it was produced by a change in preemphasis. This suggests that the warnings of Makhoul and Wolf (1972) and Fitch (1989) about the impact of tilt on LPC estimates of F1, although basically right, need to be tempered by a realization that a change of tilt produced at the glottis is not the same as a change of tilt produced by preemphasis.

---

[3] This analysis used the same parameters as the LPC analysis of the F0 items. 14 coefficients, Hamming window, window length of 258 samples at 10 kHz sampling rate.

[4] Spectral tilt was measured by fitting a regression line to the LPC spectrum.

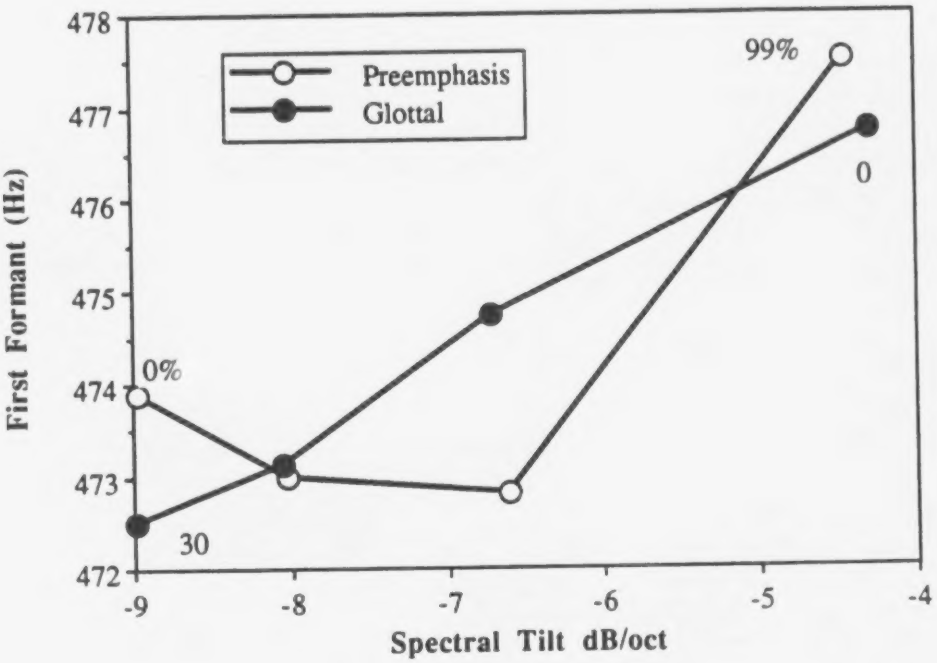


Figure 3. Two sources of spectral tilt and their effects on estimated F1.

Conclusions

The first study reported here indicates that error in the LPC estimate of F1 increases as F0 increases, but the error is non-monotonically related to F0 (the LPC estimate is both higher and lower than the actual F1 depending on the relationship between F0 and F1). The practical upshot of this finding is that the LPC error constitutes statistical noise in acoustic-phonetic studies. The magnitude of the error may preclude the use of LPC for the study of children's voices, but otherwise does not pose too serious a problem. The fact that the error could be reduced by the use of a pitch-synchronous analysis also suggests that phonetic research could benefit by the availability and use of systems such as Krishnamurthy's two-channel analysis.

The second study shows that the effect of glottal tilt on LPC estimates of F1 is monotonic and small. Because it is monotonic this source of error cannot be treated as statistical noise, but because it is small (over the range of glottal tilt variation found in natural speech) it is not overly problematic. As a precaution it is advisable that those studies which report small changes in F1 across conditions also report spectral tilt data to insure that the small formant differences are not the result of significant differences in glottal tilt.

References

Assmann, P.F. & Nearey, T. M. (1987). Perception of front vowels: The role of harmonics in the first formant region, *Journal of the Acoustical Society of America*, **81**, 520-534.

Atal, B.S. & Schroeder, M.R. (1974). Recent advances in predictive coding - Applications to speech synthesis, *Speech communication seminar*, Stockholm.

Darwin, C.J. & Gardner, R.B. (1985). Which harmonics contribute to the estimation of first formant frequency?, *Speech Communication*, **4**, 231–235.

Fant, G. (1983). The voice source: Acoustic modeling, *Speech Transmission Laboratory QPSR* **4/1982**, pp. 28–48, Royal Institute of Technology, Stockholm, Sweden.

Fitch, H. (1989). Comments on 'Effects of noise on speech production: Acoustic and perceptual analyses' [J.Acoust.Soc.Am.84, 917-928(1988)], *Journal of the Acoustical Society of America*, **86**, 2017-2019.

Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America*, **67**, 971–995.

Klatt, D.H. & Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820–857.

Krishnamurthy, A.K. & Childers, D.G. (1986). Two-channel speech analysis, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-34**, 730–743.

Makhoul, J.I. & Wolf, J.J. (1972). *Linear prediction and the spectral analysis of speech.* (Bolt Beranek and Newman Inc., Cambridge, MA), NTIS AD-749066, Rep. 2304.

Markel, J.D. & Gray, A.H. (1976). *Linear prediction of speech.* (Springer-Verlag, New York).

Miyoshi, Y., Yamato, K., Mizoguchi, R., Yanagida, M. & Kakusho, O. (1987). Analysis of speech signals of short pitch period by a sample-selective linear prediction, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-35**, 1233–1240.

Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels, *Journal of the Acoustical Society of America*, **53**, 1632–1645.

Ryalls, J.H. & Lieberman, P. (1982). Fundamental and frequency and vowel perception, *Journal of the Acoustical Society of America*, **72**, 1631–1634.

Summers, W.V., Johnson, K., Pisoni, D.B. & Bernacki, R.H. (1989). Addendum to 'Effects of noise on speech production: Acoustic and perceptual analyses' JASA 84:917-928, *Journal of the Acoustical Society of America*, **86**, 1717–1721.

Summers, W.V., Pisoni, D.B., Bernacki. R.H., Pedlow, R.I. & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and percetual analyses, *Journal of the Acoustical Society of America*, **84**, 917–928.

Wong, D.Y., Hsiao, C.C. & Markel, J.D. (1980). Spectral mismatch due to preemphasis in LPC analysis/synthesis, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-28**, 263-264.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Stress Clash in Isolated Phrases and Sentence Contexts[1]

Keith Johnson and Michael S. Cluff

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] An earlier version of this paper was presented at the 117th meeting of the Acoustical Society of America in Syracuse, New York. We appreciate the suggestions and comments of Mary Beckman and Stuart Davis. This research was supported by NIH Training Grant NS-07134-09 to Indiana University.

Abstract

The experiment reported in this paper investigates the possibility that stress clash should be called "accent clash". Stress clash in the phonological literature has been defined in terms of lexically specified stresses, while experimental investigation of stress clash candidates defined in this way has failed to find evidence for such rhythmic effects in speech production. This study tests the hypothesis that lexical stress defines the candidates for a clash and that the placement of accents during speech production determines whether a clash will actually occur. Subjects produced pairs of adjective noun phrases in which the presence of a stress clash was manipulated (loose cannon/ loose canoe) both as isolated phrases and within sentence environments. It was found that subjects showed the segmental duration effects which are predicted by metrical phonology when the clashing stressed syllables were both accented (the phrase contexts) but not when only the noun was accented (the sentence contexts). It was also found that the temporal location of the peak of F1 was not affected by the clash/nonclash manipulation and that F1 trajectories for the clash and nonclash conditions diverged after the peak of F1.

Stress Clash in Isolated Phrases and Sentence Contexts

Patterns of segmental duration variation can be attributed to a number of segmental and structural factors. Among the factors commonly acknowledged are (1) inherent segment duration, (2) segmental context, (3) lexical stress, (4) location of a word within a syntactic structure, and (5) the presence or absence of emphasis. Klatt (1976) demonstrated that most of the observed phenomena in the literature on segmental duration in American English can be accounted for if these factors are included in a concatenative model of segment duration.

In contrast to a concatenative model of segmental duration, a hierarchical model has been proposed as appropriate for the description of rhythmic properties of speech production (Pike, 1943; Lehiste, 1977; Huggins, 1975,1978; Fourakis and Monahan, 1988), and thus, indirectly also for some aspects of segmental duration. According to this hypothesis, in addition to the segmental and structural factors which play a role in concatenative models, rhythmic structure affects segmental durations in speech production.

Although this metrical approach is intuitively appealing, there is very little experimental evidence for rhythmic effects in speech production. It has been clearly demonstrated that, even in the most favorable situations, English speakers do not produce speech in which the intervals between stressed syllables are isochronous (see for example Nakatani et al., 1981). Rather, the strongest evidence that rhythm plays a role in speech comes from studies of speech perception (see Lehiste, 1977; Darwin and Donovan, 1980; Martin, 1970; and Allen, 1975). The evidence against isochrony in speech production taken together with the evidence that rhythm plays a role in speech perception indicates that rhythm (if it plays a role in speech production at all) is only one of a number of factors which determine segmental durations. This further suggests that evidence of rhythmic effects in speech production will be in the form of subtle (but predictable) changes in segmental durations within the constraints of other, concatenative factors affecting segmental duration.

Cooper and Eady (1986) took this attitude and measured utterances for which the metrical theory of phonology (Liberman and Prince, 1977; Selkirk, 1984) makes explicit predictions concerning rhythmic effects in speech production. They investigated two putative phenomena, (1) stress shift and (2) stress clash. In stress shift, it is predicted that the location of stress in a multi-syllabic word will be shifted to the left (earlier in the word) in order to preserve an alternating stress pattern. For example, "bamboo" is normally stressed on the last syllable (e.g. bamBOO), but in the phrase "bamboo tables" the stress seems to fall on the first syllable (BAMboo TABles). This stress shift results in an alternating stress pattern which is the prefered state of affairs in English and perhaps universally (Selkirk, 1984, p. 12, *passim*). Stress clash has a similar description and motivation, except that in stress clash situations it is not actually possible to shift the location of stress within the word because there is no "stressable" syllable to the left of the lexically stressed syllable and thus, other rhythmically motivated effects occur. For example, the phrase "cement tables" involves the same type of situation which results in stress shift in "bamboo tables", but the reduced vowel in the first syllable of "cement" (in some dialects of English) is unstressable. Of course, when there is only one syllable in a word, stress clash is the only possibility (e.g. "big tables"). In cases of this type, Selkirk (1984, p. 186ff) has suggested that the first stressed syllable (the "-ment" of "cement" or "big") is phonetically lengthened "as a manifestation of the tendency toward isochrony of beats." Cooper and Eady (1986) tested both the stress shift and stress clash predictions of (this version of) metrical phonology and found evidence for neither. So, even with rather more subtle expectations concerning the phonetic realization of speech rhythm, there is still very little evidence for rhythmic patterns in speech production.

Johnson and Evans (1987) replicated one of the experiments conducted by Cooper and Eady (1986) and added a manipulation of speaking style. Subjects read sentences in which the presence of a stress clash was manipulated in both a normal reading style and in a careful reading style (as if speaking over a bad telephone line). In neither reading style was there evidence for a durational adjustment resulting from a stress clash. Johnson and Evans suggested that rhythmic phenomena such as stress clash may depend less on lexically specified stressed syllables and more on accent placement within an utterance. This relates back to Bolinger's (1972) maxim, "Stress belongs to the lexicon. Accent belongs to the utterance." Lexically specified stressed syllables may or may not be accented when an utterance is pronounced, because the presence of a pitch prominence (intonational accent) on a particular word depends on semantic, pragmatic and perhaps syntactic factors (Bolinger, 1972; Bresnan, 1972; Schmerling, 1976; Selkirk, 1984). The hypothesis of Johnson and Evans (1987), which was tested in the experiment reported here, is that the "beats" in speech rhythm are the intonationally prominent syllables, not lexically stressed syllables. In order to test this hypothesis we had subjects read utterances which involved stress clash and nonclash environments (as defined by lexically specified stress) in two conditions. In one condition, only one of the lexically stressed syllables was given a pitch prominence; in the other condition, both lexically stressed syllables received a pitch prominence. We predicted that there would be a durational adjustment to stress clash only in those utterances which also involve an accent clash. The data of this experiment suggest that stress clash is best defined at the level of the utterance and in terms of pitch accents rather than at the level of lexical stress. "Stress clash" is therefore really "accent clash".

Method

Materials. In Webster's Pocket Dictionary , all nouns with primary lexical stress on the first or second syllable were identified through the use of a lexical search program. Sixty-five noun pairs which had the characteristics shown in Table 1(a) were selected. Examples of noun pairs that fit these criteria are: person–percent, broker–brochure, raven–ravine.

Insert Table 1 about here

Next, a semantically appropriate adjective was selected for each noun pair, and the nouns were paired with both the adjective and its comparative or superlative (two syllable) counterpart. Examples of the test noun phrases are shown in Table 1(b). The phrases were embedded as the first NP in simple NP-VP sentences. In addition, the number of syllables in the sentences were balanced, in order to control for possible effects of sentence length. Thus, 260 sentences were constructed, four for each noun pair. Examples of the sentences are shown in Table 1(c). Note that the first sentence of each pair with a monosyllabic adjective contains a lexical stress clash, while the second does not.

Procedure. Two native speakers of American English, one male and one female, who were unfamiliar with the purpose of the study, spoke each of the 260 items twice. In the first session, subjects read the test noun phrases in their sentence contexts. The subjects were instructed to emphasize the noun of the subject noun phrase in a normal declarative intonation. In the second session, the noun phrases were produced in isolation. The subjects were instructed to read the phrases as if someone had just made a statement involving the item in the phrase, and that they were repeating the phrase with some surprise. The subjects produced phrases which had two pitch accents as illustrated in Figure 1. In this figure, the average of the F0 contours of the items produced in *phrases* (by subject LM) have two pitch prominences, while the average F0 contour of the items produced in *sentences* is characterized by only one pitch accent.[1]

Items to be read were presented to subjects on a video terminal located in a sound-attenuated booth and item presentation, randomization and digital sampling were all under computer control (Dedina, 1987).

Insert Figure 1 about here

Thus, there were three independent variables manipulated in this study; (1) clash versus nonclash environments (such as "loose cannon" versus "loose canoe"), (2) the number of syllables in the adjective ("loose canoe" versus "looser canoe"), and (3) the number of intonational accents in the subjects' productions of each item (items read in sentences were produced with one intonational accent, while items read as isolated phrases were produced

---

[1] The contours can be transcribed in Pierrehumbert's (1980) system of intonational transcription as L+H* L* H H% (phrase condition) and H* L L% (sentence condition). Traces for Figure 1 are time normalized averages of voiced frames across all tokens.

Table 1

Criteria used in selecting materials, and examples of the test noun phrases and sentences used in the experiment.

(a) Criteria.

1. Members of the pair had the same number of syllables.
2. They had segmentally similar first syllables.
3. One member of the pair receives primary stress on the first syllable, while the other receives primary stress on the second syllable.

(b) Example Noun Phrases.

	First syllable stress	Second syllable stress
One syllable Adjective	"the loose cannon"	"the loose canoe"
Two syllable Adjective	"the looser cannon"	"the looser canoe"

(c) Example Sentences.

The loose/looser CANNON knows government secrets. The loose/looser CANOE glides along the current.
The large/larger BUFFER fills the computer's memory. The large/larger BUFFET fills a very long table.
The long/longer OVERTURE precedes a short opera. The long/longer OVATION inspires the singer.

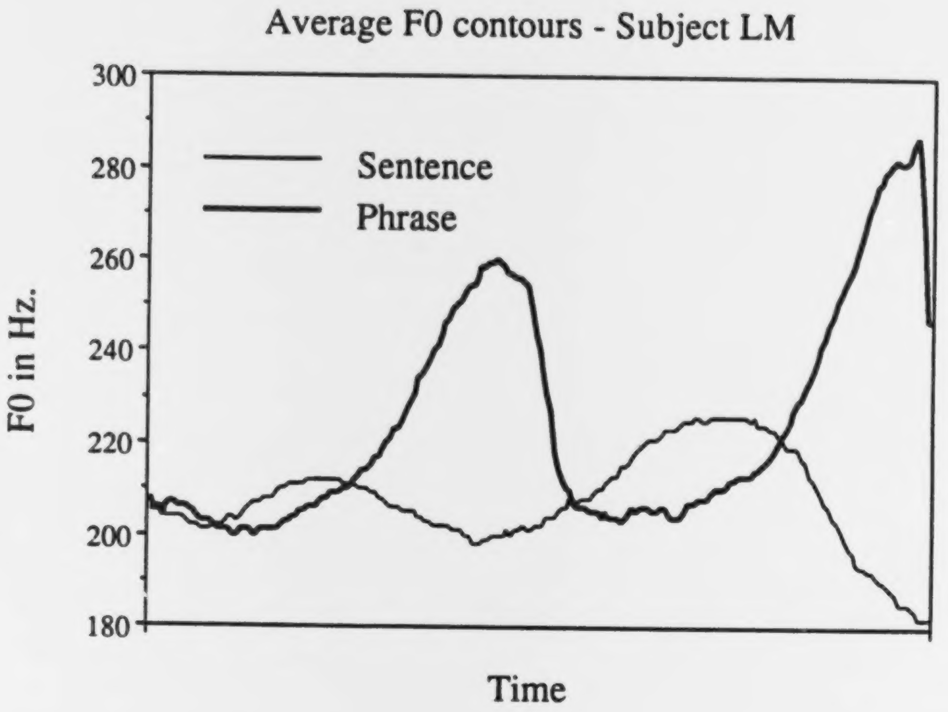


Figure 1. Average F0 contours of the test NP's for the items produced in sentence contexts and items produced in isolated phrases (subject LM). The phrase items show pitch prominence for both the adjective and the noun.

with two pitch accents). The dependent variables were the duration of the adjectives and the temporal locations of the peaks of F1 and RMS amplitude during the first syllable of the adjective.

Results

Durations of the adjectives were measured from waveform and pseudo-spectrogram displays of the digital signal.[2] F1 was estimated by LPC analysis at intervals of 12.8 ms throughout the adjective. The autocorrelation method of LPC was used to calculate LPC coefficients and a peak-picking algorithm was used to find F1. Gross errors and missing values were corrected by interpolation and the F1 contours were then smoothed by a three point smoothing filter.

Insert Figure 2 about here

The adjective duration results are shown in Figure 2. There was a main effect for sentence versus phrase contexts. Adjectives produced in phrases were longer than those read in sentences $[F(1,128) = 2064.1, p < 0.001]$ (phrase=469.6 ms, sentence=338.2 ms). We may note two possible sources of this effect. First, the adjectives produced in phrases carried an intonational accent and thus, may be considered more emphatic, or focused than those same adjectives in sentential context (see Klatt, 1976 on the effect of emphasis on segmental duration). Second, Huggins (1978) has suggested that, "The more words there are in a sentence, the shorter each word tends to become" (p. 287). Without addressing the issue of whether the unit of analysis for this generalization should be the sentence or the intonational phrase, we may point out that the extra length for adjectives produced in phrases fits Huggins' observation.

The statistical analysis also revealed (unsurprisingly) an effect of the number of syllables on overall word duration $[F(1,128) = 383.3, p < 0.001]$. Two syllable adjectives were longer than one syllable adjectives (448.6 ms versus 359.3 ms).

There was a main effect for the stress clash factor $[F(1,128) = 21.04, p < 0.001]$ (nonclash=398.6 ms, clash=409.2 ms) as well as a significant interaction of stress clash and accentual structure (phrase context versus sentence context) $[F(1,128) = 19.41, p < 0.001]$.

---

[2] We also measured the interval between the onset of the adjective and the onset of the following noun (i.e. adjective plus pause). An analysis of the onset-to-onset data was also conducted, but because this analysis showed the same statistically reliable effects which were found in the analysis of the adjective duration data, only one set of data is presented here (the adjective duration data).

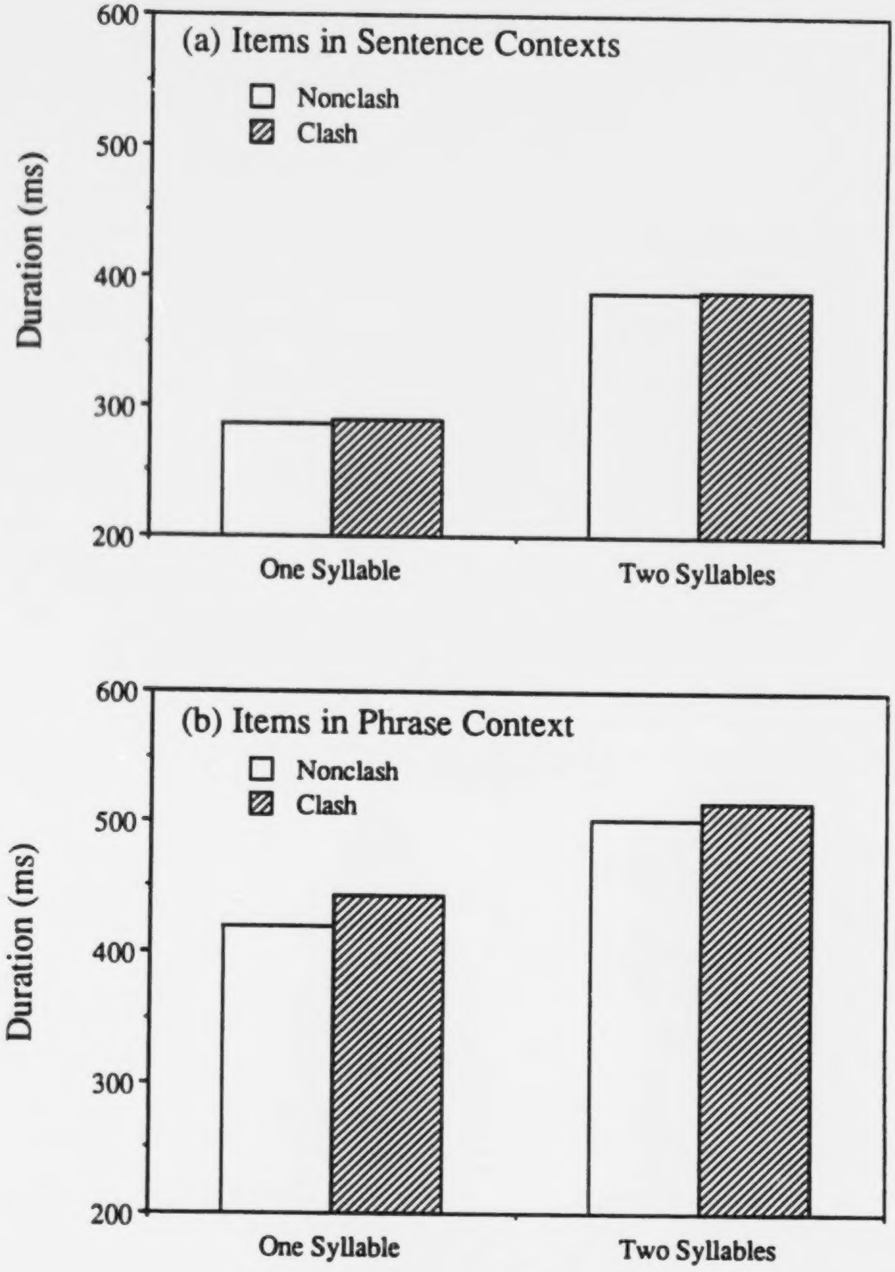


Figure 2. Adjective duration results averaged across subjects. (a) Items produced in sentence contexts. (b) Items produced in phrase contexts. The filled bars are the clash tokens and the open bars are the items produced in no-clash environments.

This interaction is shown in Figure 2. In a post-hoc comparison of means it was found that the difference between stress clash and no stress clash was reliable only for the items which had been produced in the phrase condition (for both monosyllabic and bisyllabic adjectives). This was true for both subjects. The fact that there was no effect of stress clash in the sentence condition replicates the findings of Cooper and Eady (1986), while the change in the duration of the adjective in the phrase condition supports our view of speech rhythm.

There was also an interaction between the accentual structure condition and the number of syllables [$F(1,128) = 26.01, p < 0.0001$]. The difference between one and two syllable adjectives was smaller when the items were produced in a phrase context. Recall that items produced in phrasal context had a pitch accent on the adjective, while items produced in sentences did not. This interaction suggests that when accent is placed on a monosyllabic word, the relative overall increase in duration is greater than when accent is placed on a bisyllabic word.

An analysis of the temporal locations of the peak of F1 in these utterances revealed no differences in the location of F1 peaks as a function of stress clash. The only statistically reliable effect in the analysis of the F1 peak data was the accentual structure main effect [$F(1,115) = 29.09, p < 0.0001$]. The peak of F1 occured on average 178 ms into the word when items were produced in sentence contexts and 189 ms into the word when items were produced in phrase contexts. This effect is consistent with the large duration difference found for these contexts, although interestingly, the magnitude of peak shift is much smaller (11 ms) than the magnitude of duration difference (>130 ms).

Similarly, only the accentual structure effect was reliable in the analysis of the RMS peak data [$F(1,128) = 110.12, p < 0.0001$]. The average location for the peak of RMS amplitude was 162.6 ms in tokens produced in sentences while the peak occurred later (210.3 ms) when the items were produced in phrases.

Figure 3 demonstrates the relevance of the peak data. This figure shows average F1 contours for subject LM's productions of monosyllabic adjectives in the phrase condition. The temporal locations of the peaks were not reliably different across the clash and noclash environments while the clash items had longer duration. As illustrated in Figure 3 the durational difference between clash and nonclash items was realized on the end of the word.

Insert Figure 3 about here

Conclusions

Overall this pattern of results suggests: (1) that rhythmic effects in speech production are best described in terms of accents rather than lexically stressed syllables, and (2) that

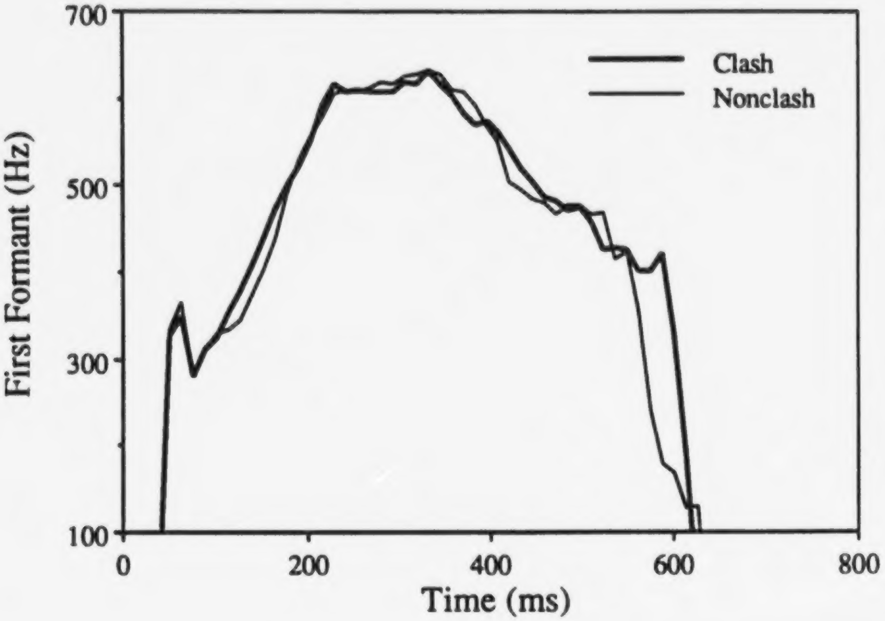


Figure 3. Average F1 contours for the monosyllabic adjectives produced in phrases by subject LM.

the rhythmic adjustment to stress clash involves a change in speech production localized on the final portion of the word and not a wholistic rescaling of articulation.

The first conclusion is motivated by the fact that we found the durational effect which is predicted by a metrical or hierarchical approach to the description of speech production, but that this effect was only found when subjects produced the utterances which involved a stress clash with accents on the "clashing" syllables. In other words, we found a durational adjustment to accent clash, but not to stress clash. Our null result in the case of stress clash is consistent with that reported by Cooper and Eady (1986) and so the two studies are in that sense mutally confirming. However, our results suggest a different conclusion. Cooper and Eady concluded, "At least some of the presumed 'facts' of rhythmic patterns presented in metrical phonology do not hold up under empirical testing" (p. 383). We would rather conclude that the rhythmic patterns described in metrical phonology have been wrongly attributed to lexically determined stresses when, in actuality, they are better described as properties of actual (rather than potential) pronunciations; the timing of pitch accents.

Insert Figure 4 about here

The second conclusion is motivated by the fact that the temporal locations of the peaks of F1 and RMS amplitude were not reliably affected by stress clash even in those cases in which overall duration was affected. This indicates that the durational difference between clash and no-clash environments occurs over the last part of the word. Note that we found the same pattern of results (durational difference coupled with no change in peak of F1 or RMS amplitude) in both monosyllabic and bisyllabic adjectives. This suggests that intervening syllables (whether one or two) impinge upon the closing gesture of an accented syllable and have almost no effect on the opening gesture. This is illustrated in Figure 4. The top panel of this figure shows hypothetical F1 contours of the four types of utterance used in this experiment. In these idealized schema the peaks of sonority for the clash and no-clash items are isochronous. In the nonclash case the intervening unstressed syllable(s) overlaps with the closing gesture of the first word and the opening gesture of the second word. The bottom panel in Figure 4 is data from subject EG. Each function is an average across the 65 tokens produced in each condition (items produced in phrases). The data diverge from the hypothetical situation in that the interval between the F1 peaks for the first and second words are clearly dependent upon the number of intervening syllables (anisochrony). They are, however, similar to the hypothetical data in that the closing portion of the first word in the clash condition is different from that in the no-clash condition. This data (for one syllable adjectives, at least) is comparable to that reported by Beckman (1989), "the jaw opening gesture was relatively shorter and the closing gesture relatively longer in the stress clash context."

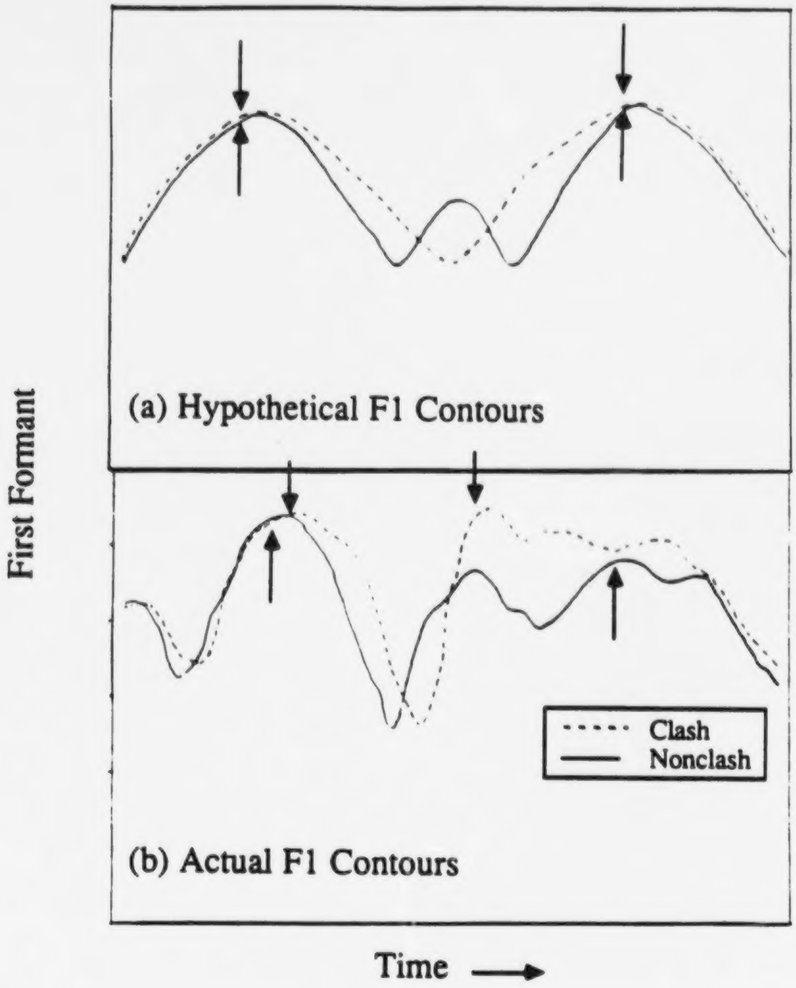


Figure 4. Hypothetical (a) and actual (b) F1 contours. The actual contours are averages (n=65) of subject EG's productions of the monosyllabic and bisyllabic adjectives (F1 traces of the entire noun phrase are shown) in clash and no-clash conditions in phrases. The contours were aligned at the onset of the phrase.

References

Allen, G. D. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, **3**, 75–86.

Beckman, M. E. (1989). Articulatory correlates of stress clash rhythms. *Journal of the Acoustical Society of America*, **85**, S96.

Bolinger, D. (1972). Accent is predictable (if you're a mind-reader). *Language*, **48**, 633–644.

Bresnan, J. (1972). Stress and syntax: A reply. *Language*, **48**, 326–342.

Cooper, W. E. & Eady, S. J. (1986). Metrical phonology in speech production. *Journal of Memory and Language*, **25**, 369–384.

Darwin, C. J. & Donovan, A. (1980). Perceptual studies of speech rhythm: Isochrony and intonation. In Simon, J., editor, *Spoken language generation and understanding*. D. Reidel, Dordrecht.

Dedina, M. (1987). SAP: A speech acquisition program for the SRL-VAX. *Research on speech perception progress report no.13*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana Universiy.

Fourakis, M. & Monahan, C. B. (1988). Effects of metrical foot structure on syllable timing. *Language and Speech*, **31**, 283–306.

Huggins, A. (1975). On isochrony and syntax. In Fant, G. & Tatham, M., editors, *Auditory analysis and perception of speech*. Academic Press, New York.

Huggins, A. W. F. (1978). Speech timing and intelligibility. In Requin, J., editor, *Attention and performance VII*. Lawrence Erlbaum Associates, Hillsdale, N.J..

Johnson, K. & Evans, D. (1987). Phonetic reality and phonological prediction: Stress clash and rhythm in English. *Journal of the Acoustical Society of America*, **81**, S66.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208–1221.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, **5**, 253–263.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, **8**, 249–336.

Martin, J. G. (1970). Rhythm-induced judgements of word stress in sentences. *Journal of Verbal Learning and Verbal Behavior*, **9**, 627–633.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of american english speech rhythm. *Phonetica*, **38**, 84–106.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. *Unpublished Ph.D. dissertation, MIT.*

Pike, K. L. (1943). *Phonetics.* The University of Michigan Press, Ann Arbor, Michigan.

Schmerling, S. (1976). *Aspects of English sentence stress.* University of Texas Press, Austin, Texas.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and stucture.* MIT Press, Cambridge, MA.

Final Report to the NTSB
on the Speech Produced by the Captain of the Exxon Valdez[1]

Keith Johnson, David B. Pisoni and Robert H. Bernacki

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The analyses reported in this paper were carried out in connection with the National Transportation Safety Board (NTSB) investigation of the Exxon Valdez accident that occurred on March 24, 1989.

Abstract

In this report we consider the possibility that speech analysis techniques may be used to determine whether an individual was intoxicated at the time that a voice recording was made, and discuss an analysis of the speech produced by the Captain of the Exxon Valdez recorded at several points around the time of the accident at Prince William Sound, Alaska. A review of previous research on the effects of alcohol and other effects on speech production suggests that it may be possible to attribute a certain, unique pattern of changes in speech to the influence of alcohol. However, the rate of occurrence of this pattern or the reliability of a decision based on observations such as these is not known. Acoustic-phonetic changes observed in a small number of tokens of Captain Hazelwood's speech recorded before, during and after the accident revealed a number of changes in speech behavior which correlate well with the findings of previous research on the effects of alcohol on speech production.

Final Report to the NTSB on the Speech Produced by the Captain of the Exxon Valdez

In this report, we briefly summarize previous research on the effects of alcohol on speech production and previous research on other effects on speech production. We then discuss an analysis of the speech produced by the Captain of the Exxon Valdez recorded at several times before, during and after the accident at Prince William Sound.[1]

The Problem of Unique Specification

Before discussing this particular case, we wish to place the present investigation within a general framework. The question which we are implicitly attempting to address in this report is whether it is possible to determine if an individual was intoxicated at a particular point in time based on acoustic analyses of voice recordings. This question hinges crucially on whether there are properties of speech which occur when a speaker (any speaker) is intoxicated and which do not occur in any other circumstance. We will call this the problem of unique specification.

In the following section, we review several studies which have found that there are a number of acoustic-phonetic characteristics of speech which occur when individuals are intoxicated. This research is an important first step in determining whether speech patterns may uniquely specify alcohol intoxication, but, to our knowledge, there is no published research which directly addresses the problem of unique specification. In spite of this lack of previous research, there are at least two reasons to believe that voice recordings may contain reliable information which uniquely indicates that an individual was intoxicated at the time of the recording. These have to do with the physiological and pharmacological effects of alcohol and the complexity of speech motor control.

Although the effects of alcohol at a cellular level in the nervous system are not fully understood[2], the general functional effects are clear. "The principal effects of acute dosage of ethyl alcohol are observed in the nervous system, where there is a progressive and simultaneous impairment of function at many levels" (Berry & Pentreath, 1980, p. 43). Ethanol diffuses easily through cell boundaries (Wallgren & Barry, 1970, p. 36), and results in a biphasic neural response. At low concentrations, nerve cell excitability is increased, while at high concentrations there is a progressive reduction of excitability (p. 254). This reduction in nerve cell excitability leads to behavioral responses to alcohol which (particularly relevant for speech) include decreased motor coordination.

---

[1] The tapes that we analyzed and information concerning the communications/recording equipment, the times of the recordings and the results of the blood alcohol test were provided to us by the staff of the National Transportation Safety Board.

[2] Berry & Pentreath (1980) review some of the data having to do with the effects of alcohol on neural membrane permeability and the synthesis and release of neurotransmitter. They note a variety of specific cellular effects and affected sites in the nervous system.

In addition to the neurological effects of alcohol when it reaches the brain through the blood stream, it is likely that the local contact of alcohol with the surfaces of the mouth and throat have some effect on speech production. It is well known that local concentrations of alcohol in the stomach irritate the mucosa and paralyze the muscles of the stomach wall (Wallgren & Barry, 1970, pp. 40, 61). There is also some evidence which suggests that alcohol applied to the tongue (at least the tongues of cats) can produce a biphasic sensitivity to mechanical stimulation (Hellekant, 1965). These local effects of alcohol in the mouth and throat may result in effects on speech production which differ from the effects which result from other central nervous system depressants or other factors, although we are aware of no previous research which has attempted to test this hypothesis.

Tests of motor coordination (such as walking a straight line or standing on one foot with eyes closed) are commonly used to indicate whether a person is intoxicated. Speech production is another complex motor activity which requires a high degree of coordination and so may also be affected by alcohol consumption. Two types of motor complexity in speech production can be distiguished. First, speech production requires very precise inter-gestural coordination. For example, the main difference between /d/ and /t/ in English is the timing of a gesture of the vocal folds relative to a gesture performed by the tip of the tongue. The relative timing of these two gestures ("voice onset time") is measured in milliseconds (ms) (Lisker and Abramson, 1964). The onset of voicing (vocal cord vibration) for /d/ in word initial position occurs approximately simultaneously with the release of oral stop closure, while the onset of voicing for /t/ occurs 40 to 60 ms after the release of oral stop closure. Mistiming the two gestures by as little as 20 ms results in a perceptually different consonant. Second, speech involves fine motor control in moving the articulators to the target positions for different speech sounds. For example, the fricative /s/ is produced by pressing the sides of the tongue against the upper molars and depressing the center of the tongue, creating a narrow groove with the tip of the tongue. The articulatory difference between /s/ and /sh/ is very subtle even though the acoustic difference is quite large. The location of the tongue relative to the front teeth and the length of the constriction at the roof of the mouth (the tongue groove) distinguish these two sounds in speech production (Subtelny, Oya & Subtelny, 1972). If the tongue tip is kept close to the front teeth and the constriction at the roof of the mouth is relatively short (2.5 cm), an /s/ is produced. However, if the constriction is slightly longer or wider, or the tongue tip is held a little further back in the mouth, the resultant sound is more like /sh/. These observations suggest that small variations in speech timing or gestures can have acoustically reliable consequences for speech production (Stevens, 1972). Alcohol's effects on the central nervous system and the local effects of alcohol on the muscles and proprioceptors of the vocal aparatus, coupled with the inherent complexity of speech production, suggest that there may be patterns of speech production which are uniquely attributable to alcohol intoxication.

Previous Findings on Alcohol Impaired Speech

This section is a brief review of previous research on the effects of alcohol on speech production. For more complete reviews of the literature see Pisoni, Hathaway and Yuchtman (1986), Klingholz, Penning and Liebhardt (1988) and Pisoni and Martin (1989). The effects of alcohol on speech production that have been observed in controlled laboratory studies can be divided into three types: gross effects, segmental effects and suprasegmental effects. Examples of each of these effects are listed in Table 1.

Insert Table 1 about here

Gross effects involve word level alterations in speech production. These effects are very noticeable when intoxicated subjects are instructed to read a passage. Subjects may revise, omit or interject words (Sobell and Sobell, 1972; Sobell, Sobell and Coleman, 1982). It has been assumed that this class of errors reflects changes or modifications in speech planning. As neural function is depressed by alcohol, the speaker's ability to control the articulators is impaired which in turn may affect the planning stage in speech production. Thus, word level alterations occur when the subject is required to read a passage. In spontaneous speech, however, it is much harder to decide what should count as a gross error because the speaker's intended utterance is not known. Therefore, gross effects are less valuable for the evaluation of spontaneous speech and diagnosis of any impairment due to alcohol.

Segmental effects involve the misarticulation of specific speech sounds. The segmental effects which have been most often reported are: misarticulation of /r/ and /l/, misproduction of /s/ (more like /sh/), final devoicing of obstruents, and deaffrication. Examples of the last two effects are given in Table 1. Obstruent devoicing involves a problem of timing and glottal control similar to the example of /d/ and /t/ given in the previous section. The other segmental effects involve the control of the tip of the tongue. Lester & Skousen (1974) found that segmental effects such as these did not appear until subjects had consumed about 10 ounces of 86 proof straight bourbon over a period of about 3 and 1/2 hours.

Phonetic theory makes some predictions about the changes/modifications of speech articulation after alcohol consumption. These predictions derive from the study of articulatory ease (see for example Lindblom, 1983) which suggests that not all speech sounds are equally easy to produce. Evidence of this comes from studies of the development of speech in children (de Villiers & de Villiers, 1978), the patterns of historical language change (Antilla, 1972), and patterns of language dissolution in aphasia (Jakobson, 1941), as well as model studies of articulation (Lindblom, 1983). Most of the segmental effects observed in speech produced while intoxicated have analogs in these data. For instance, it is common for children to misarticulate /r/ and /l/ as in the production of "train" as /twen/. Also, final devoicing

Table 1

Summary of previous research on the effects of alcohol on speech production.

Gross effects	word/phrase/syllable interjections[1] word omissions[1,4] word revisions[1] broken suffixes[1]
Segmental effects	misarticulation of /r/ and /l/[4,5] /s/ becomes /sh/[3,4] final devoicing (e.g. /iz/ – > /is/)[3,5] deaffrication (e.g. 'church' – > 'shursh'[3,4,5])
Suprasegmental effects	reduced speaking rate[1,2,3,5] decreased amplitude[2] increase of unvoiced to voiced ratio[3,5,6] decreased spectral tilt[6] mean change in pitch range (talker dependent)[4,5,6,7] increase in pitch variability[5,6]

---

[1] Sobell & Sobell (1972). 16 alcoholics, 5-10 ounces, 86 proof alcohol.

[2] Sobell, Sobell & Coleman (1982). 16 talkers, $0.05 < BAL < 0.1\%$.

[3] Lester & Skousen (1974). Number of talkers not mentioned, 86 proof straight bourbon, one ounce/20 min. up to 14 ounces.

[4] Trojan & Kryspin-Exner (1968). 3 talkers, 1 to 1.38 liters of heavy Austrian wine (13% alcohol).

[5] Pisoni, Hathaway & Yuchtman (1986) and Pisoni & Martin (1989). 5 talkers, $0.1 < BAL < 0.17\%$.

[6] Klingholz, Penning & Liebhardt (1988). 16 talkers, $0.067 < BAL < 0.16\%$.

[7] Dunker & Schlosshauer (1964). 1 talker, "consuming alcoholic beverages liberally" and shouting.

and deaffrication are very common in child speech and in historical language development. The substitution of /sh/ for /s/, however, is not typically found in child speech, and /s/ is more common than /sh/ in the languages of the world. Therefore, this segmental effect, rather than being the result of a general loss of motor coordination (as is most likely the case for the other segmental effects), seems to have a different cause. The change of /s/ to /sh/ may be related to loss of responsiveness of the surface muscles of the tongue or a loss of proprioceptive feedback from the tongue after direct contact with ethanol during consumption.

Suprasegmental effects are perhaps more perceptually salient than segmental effects, but require quantification. These effects involve the rate and amplitude of speech and vocal cord function. Trojan & Kryspin-Exner (1968) reported an increase in voice fundamental frequency (rate of vocal cord vibration). Pisoni & Martin (1989) found that fundamental frequency decreased for some, but not all subjects. Klingholz et al. (1988) also found a tendency for decreased fundamental frequency. Fundamental frequency (F0) is also more variable in speech produced while intoxicated when compared to a control condition (Pisoni & Martin, 1989; Klingholz et al., 1988). Klingholz et al. (1988) also found that the speech harmonics-to-noise ratio decreased after alcohol intoxication. This measure reflects a change in the mode of vocal cord vibration indicative of increased breathiness after alcohol intoxication. They also found a change in the long-term average (LTA) spectrum in intoxicated speech. There was an increase in high frequency energy, which may reflect an increase in the unvoiced/voiced ratio after alcohol consumption (as reported by Pisoni & Martin, 1989). All of these effects can be measured directly using digital signal processing techniques (see Pisoni & Martin, 1989 and Klingholz et al., 1988).

The effects on speaking rate and F0 can be related to the general physiological effects of alcohol in the following ways. The reduction in speaking rate may be the result of an attempt to compensate for the loss of motor coordination which accompanies intoxication. The effect of alcohol on F0 seems to have an origin in the interaction of alcohol and the tissue of the vocal cords. Klingholz et al. (1988) suggest that the effect of alcohol on F0 may be the result of irritation and swelling of the mucous membranes of the vocal cords and desensitization of the proprioceptors of the vocal cords. They cite evidence from Dunker & Schlosshauer (1964) which indicates that vocal cord vibration after alcohol consumption (like vocal cord vibration for people with hoarse voices) is more variable and lower in pitch. Klingholz et al. posited a connection between vocal cord swelling due to mechanical stress (shouting or speaking for an extended time) and swelling due to alcohol consumption. This explanation may also account for the increase in the unvoiced/voiced ratio in intoxicated speech.

Other Effects on Speech Production

In this section, we briefly review some of the previous research on environmental and emotional effects on speech production and compare these effects with the effects of alcohol on speech production. Table 2 is a summary of some previous research addressing environmental and emotional effects on speech production. As indicated in this table, most researchers who have investigated the effects of these factors on speech production have focussed on suprasegmental phenomena. Only occassionally have segmental phenomena other than vowel formant measures been investigated. This research focus reflects a practical concern for the design of automatic speech recognition devices for use in a variety of circumstances, where suprasegmental changes and some types of segmental changes could be detrimental to the performance of recognition systems. Therefore, the data-base we are reviewing here is not entirely comparable to that collected in the study of the effects of alcohol on speech.

Insert Table 2 about here

Hansen (1988) and Summers et al. (1988) studied the effects of *noise* on speech production (the Lombard effect). These studies found that speech produced with a high level of noise at the ears had increased fundamental frequency (F0) and duration, and reduced spectral tilt.[3] The spectral tilt measure indicates that there was a relative increase of high frequency glottal energy in the Lombard condition. Surprisingly, Hansen (1988) found no change in amplitude. The Summers et al. (1988) result is in better agreement with earlier research. Finally, the studies indicate some individual variability in the effect of noise on vowel formant values.

Moore & Bond (1987) studied the effects of *acceleration* and *vibration* on speech produced by two subjects. The two situations resulted in comparable effects on F0, intensity and vowel formants. F0 increased relative to that found for the same subjects in benign environments, vocal intensity was unchanged and vowels were less distinctive (more like /ə/). There was individual variability in the effect of acceleration on segmental duration, while speaking rate increased (reduced segmental durations) in the vibration condition. The small number of subjects in these studies is problematic, but this is the only available data on these environmental effects.

A large number of studies have employed *workload* tasks to simulate environments with high cognitive demands such as airplane cockpits. These studies have generally found that speech produced while performing a cognitively demanding task has higher F0, decreased spectral tilt and increased intensity. Data on the variability of F0 (SD F0) is mixed. This reflects a problem in the use of this measure due to the fact that F0 variability can be affected

---

[3] Hansen (1988) measured the tilt of the glottal spectrum (after inverse filtering) while the other authors listed in Table 2, who reported spectral tilt changes, measured changes in the spectral tilt of the unfiltered speech signal. There is general agreement between studies using the two measures, although note that valid tilt comparisons using the simpler method require careful control of the phonetic content (particularly vowel qualities) of the tokens being compared.

Table 2

Summary of some recent research on environmental and emotional effects on speech production.

Condition	F0	SD F0	Jitter	Tilt	Duration	Intensity	Formants
Noise[1]	⇑	⇑		⇓	⇑	NC	F1 ⇓
Noise[2]	⇑			⇓	⇑	⇑	F1 ↑
Acceleration[3]	⇑				⇕	NC	centralized
Vibration[3]	⇑				⇓	NC	centralized
Workload[1]	⇑	⇑		⇓	NC	↑	F1 & F2 ⇑
Workload[4]	↑	↓		⇓	↓	↑	NC
Workload[5]	⇑	↑			⇓	⇑	
Stess[6]	↑		↓		⇓	⇑	
Stress[7]	↑		⇓				
Stress[8]	⇕				⇕	↑	
Perceived Stress[8]	⇑				⇑	⇑	
Fear[9]	⇑	⇑	⇑	NC	⇑		NC
Fear[1]	⇑	⇑		⇓	NC	⇑	F1 & F2 ⇑
Anger[9]	⇑	⇑	NC	⇓	NC		F1 ⇑
Anger[1]	⇑	⇑		⇓	⇑	⇑	F1 ↑
Sorrow[9]	⇓	⇓	⇑	⇑	⇑		NC
Depressed[1]					⇑		centralized
Intoxicated[10]	⇕	⇑		⇓	⇑	↓	

⇑= reliable *increase* for all subjects.
↑= increase for some, but not all subjects.
⇓= reliable *decrease* for all subjects.
↓= decrease for some, but not all subjects.
⇕= some subjects showed a reliable increase, while some a reliable decrease.
NC = no change.

---

[1] Hansen, 1988 (8 talkers).

[2] Summers, et al., 1988, see also Pisoni, et al., 1985 (2 talkers).

[3] Moore & Bond, 1987 (2 talkers).

[4] Summers, et al., 1989 (5 talkers).

[5] Griffin & Williams, 1987 (20 talkers).

[6] Brenner & Shipp, 1988 (17 talkers).

[7] Brenner, Shipp, Doherty & Morrissey, 1985 (7 talkers).

[8] Streeter, et al., 1983 (2 talkers).

[9] Williams & Stevens, 1972, see also Williams & Stevens, 1981 (3 talkers).

[10] See Table 1.

in two very different ways. Variability will be reduced if the F0 contour of utterances are more monotonic in the workload condition (as suggested by Summers et al., 1989) or if there is less period-to-period variation in the vibratory cycle of the vocal cords (as suggested by Brenner et al., 1987, who also used a cognitively demanding task). On the other hand, F0 variability could be increased if utterances in the workload task had more extreme fluctuations in their F0 contours even if vocal cord jitter (period-to-period variation of F0) were reduced. Williams & Stevens (1972) provide a good example of the conceptual distinctions which need to be maintained in this area, although they did not have digital signal processing techniques at their disposal. They reported both changes in F0 range and (inferences about) changes in F0 jitter. In the absence of this distinction in some of the research on the effects of cognitive workload, it is impossible to determine whether the reported differences in F0 variablity in speech under workload reflect real individual differences or merely differences in data collection techniques. Table 2 also indicates some differences across studies in the effects of workload on segmental duration, although it is interesting that the study on the effects of workload which employed the greatest number of subjects (Griffin & Williams, 1987) reported a consistent decrease in duration. Finally, there is also some discrepancy concerning the effects of workload on vowel formant frequencies.

The term ***psychological stress*** has been used to describe situations ranging from lying to being in a fatal airplane crash. Scherer (1981) outlined some predictions for speech production in stressful situations based on the general physiological response to stress (similar to the discussion above of physiological predictions for the effects of alcohol) and then concluded that "virtually all of the studies in this field have found very strong individual differences in terms of the number and kind of vocal parameters that seem to accompany stress" (p. 179). He focussed on two problems in the literature, (1) the likelihood that subjects in laboratory studies of stress were differentially stressed, and (2) the fact that "subjects may differ in terms of the degree of control they can exert as far as their vocal production under emotional arousal is concerned" (p. 180).[4] In spite of these problems, some general trends emerge from studies of stress in laboratory and real-life emergency situations. These are indicated in Table 2 and include an increase in F0, an increase in intensity, and a decrease in F0 jitter. Brenner, Shipp, Doherty & Morrissey (1985) examined F0 jitter in situations of high stress by analyzing voice recordings of pilots involved in aircraft crashes. They found that speech in stressful situations had increased F0, and decreased F0 jitter. In a related laboratory study, Brenner et al. (1985) also found that the activity of the cricothyroid muscle, which is the primary muscle of the larynx involved in controlling F0, increased as stress increased. This provides an explanation of both the increased F0 and decreased F0 jitter found in the other studies.

Streeter, MacDonald, Apple, Krauss and Gallotti (1983) reported a case of individual

---

[4] Both of these problems have analogues in studies of the effects of alcohol on speech. Although, it is possible to objectively measure the subjects' blood alcohol level, not all previous research on the effects of alcohol on speech production have reported BALs. Also, subjects may differ in the degree of articulatory control they can exert while intoxicated.

variability in the vocal effects of stress. They examined a recorded telephone conversation between a system operator and chief system operator for Consolidated Edison during the New York City blackout, July, 1977. One talker had increased F0, duration, and amplitude as the situation developed (and presumably stress increased), while the other showed a different pattern (decreased F0 and duration, and no change in amplitude). This study illustrates Scherer's (1981) point about individual differences in response to stressful situations, and suggests that there may be no consistent phonetic pattern for any but the most extremely stressful, life-threatening situations. Interestingly, though, Streeter et al. found that naive listeners used phonetic cues consistently in making judgements about the degree of stress being experienced by the talker. Listeners judged utterances with higher F0, higher amplitude and longer segment durations as more stressed even though, for one speaker, these judgements were not correlated with degree of experienced stress. The speech parameters which were found in this study to be correlated with *perceived stressed* are listed in Table 2. Streeter et al. concluded that listeners have stereotyped expectations for vocal responses to stress, which evidently are accurate for the most extreme levels of stress, but speakers who are actually experiencing some less than maximal degree of stress do not always fit the perceptual stereotype.

Table 2 also presents a summary of several studies on the effects of emotional state (fear, anger and sorrow) on speech production. The study of the effects of emotion on speech production involves methodological problems that are not involved in the study of environmental effects on speech, where it is possible for the experimenter to create conditions which can be carefully controlled and described. In order to study the effects of emotion on speech production, however, it is necessary to rely on subjective measures of the emotional (mental) state of the speaker or have speakers simulate various emotions. In spite of these methodological difficulties, we are including this summary of previous research in an attempt to present a complete review of the factors that may affect speech production.

Williams & Stevens (1972, 1981) hired three actors to perform short plays in which the characters displayed various emotions. Their data are summarized in Table 2 and compared with some recent data from Hansen (1988), who studied the effect of *fear* by having his subjects read a prepared wordlist as they were decending steep drops on a roller-coaster. There is good agreement between these two studies concerning the effects of fear on F0. Both found that F0 increased and that F0 variability increased. Williams & Stevens also suggested that, in addition to increased F0 range, F0 jitter increased. Whereas Williams & Stevens reported no change in spectral tilt, Hansen found that the glottal spectrum was flatter in the fear condition. The more sophisticated signal processing techniques employed by Hansen may have allowed him to detect a small change not seen by Williams & Stevens. The two studies also found different effects on segmental duration. Hansen found no change, while Williams & Stevens found an increase in word duration of about 30 ms. This seems to reflect a real difference, and again may be a result of methodological differences. Hansen reported that intensity increased in the fear condition. This effect is consistent with findings for psychological stress and increased workload and seems to reflect a change in arousal

(Scherer, 1981). Finally, Hansen found changes in the first two vowel formants which were not found by Williams & Stevens.

Hansen (1988) and Williams & Stevens (1972) also studied the effects of *anger* on speech production. Here the two studies had similar methodologies and very similar results. They both found that F0, F0 varibility and F1 increased, and that spectral tilt decreased. Williams & Stevens found no changes in F0 jitter, although they were using a somewhat crude measure (fluctuation in narrow-band spectrograms). Hansen found an increase in intensity. The only discrepancy between the two studies has to do with the effect of anger on speaking rate. Where Williams & Stevens found no reliable change, Hansen also found that speaking rate decreased (increased segmental durations) in the anger condition. Notice the similarities between the effects of anger and the effects of workload and fear.

The final emotion listed in Table 2 is *sorrow*. Again, the data listed in the table are from Williams & Stevens (1972) and Hansen (1988).[5] Speech produced by actors protraying sorrow was characterized by decreased F0, decreased F0 range but increased F0 jitter. Williams & Stevens also found that spectral tilt increased in the sorrow condition (i.e. that there was a reduction of high freqency energy). Both Hansen and Williams & Stevens found an increase in segmental durations, but they found different effects on vowel formants. Williams & Stevens found no change in vowel formants while Hansen suggested (based on very few measurements) that vowels were more centralized in the depressed condition.

We have also included in Table 2 a summary of the suprasegmental effects found in the studies of alcohol and speech which were listed in Table 1. There are no situations or emotions listed in Table 2 which have exactly the same pattern of effects found in the studies of alcohol and speech, and so, given adequate measures of these acoustic correlates, it would be possible to classify the changes observed across two or more samples of speech as more like the pattern found for intoxicated speech than, for instance, speech produced in noise. It is not possible, however, to give any kind of confidence rating to such a classification, because there is not enough published data on individual differences which would allow the calculation of hit rates and false alarm rates for classifications based on these measures (this is true of the other effects shown in Table 1 also).

Another problem with classifying speech samples is that there are some possible physiological effects on speech production, which have not been previously studied. The effect of fatigue on speech production has not been examined in any controlled study of speech production. Also, we lack any data on speech production just after the speaker has been awakened. Our subjective impression is that speech produced in these circumstances may involve changes in vocal cord activity (extremely low F0 or pulse register phonation), decreased speaking rate and perhaps some effects related to dehydration of the mucous membranes in the mouth, which may be similar to the effects seen after alcohol consumption. However, the relevant controlled laboratory studies haven't been done. There are also no data on more complex situations involving combinations of effects. For instance, no one has studied what happens to speech when the speaker is both tired and under stress.

---

[5] The data reported by Hansen are based on a small number of observations. These data are included in the table because they come from a real life situation (recordings made during counselling sessions in a psychiatrist's office) and as such offer some degree of validation of the observations of Williams & Stevens.

The Speech of Captain Hazelwood

We have analyzed five different samples of speech provided to us by NTSB. Also, we examined a small number of utterances from Captain Hazelwood's televised interview with Connie Chung which was broadcast on March 31, 1990. We will refer to the speech samples according to the times at which they were recorded: (-33) 33 hours before the accident[6], (-1) one hour before the accident, (0) immediately after the accident, (+1) one hour after the accident, (+9) nine hours after the accident and (CC) televised interview. We will discuss gross errors, segmental changes, and suprasegmental changes.

Insert Table 3 about here

Gross Errors

Several of the speech errors in the NTSB tapes may be classified as gross phonetic errors. These are listed in Table 3. Note, however, that such phenomena are not uncommon in spontaneous speech regardless of alcohol consumption. What is needed in order to evaluate the condition of the speaker is a large amount of speech in which it is possible to compare the rate of occurrence of such errors across speech samples. Also, since the talker was not reading a prepared text, it is a matter of subjective judgement to say that something is or is not an error. To attempt to control for this problem, we are only reporting cases in which

---

[6] It is important to note here that the recording made 33 hours before the accident has a different history than the other recordings. All of the NTSB recordings were initially recorded using the same Coast Guard equipment, but this sample was then re-recorded onto a handheld cassette recorder before the original tape was mistakenly erased. The recording which we analyzed was produced by playing back the cassette tape using the same cassette recorder which had been used to record the sample. We investigated the possibility that the recording was corrupted by analyzing an unidentified background sound which seemed to be present in both the -33 sample and in the -1 sample. In the -33 recording, the sound had a higher average fundamental frequency (480 Hz, $n=4$ versus 472 Hz, $n=10$) and a greater F0 range (438 Hz to 588 Hz versus 456 Hz to 481 Hz) as compared with the -1 recording. The variability of the F0 in the -1 recording suggests that the sound was not constant in frequency and, thus, is not an adequate benchmark for determining the validity of the -33 recording. However, even if the -33 recording is corrupted by tape speed fluctuations of the magnitude indicated by these measurements (-9% to +22%), this degree of difference is not enough to account for the changes in speech production we report below.

Table 3

Summary of phenomena found in the analysis of the NTSB tape. Numbers in parentheses indicate the time of recording.

Gross effects	revisions (-1) Exxon Ba, uh Exxon Valdez (-1) departed – > disembarked (-1) I, we'll (-1) columbia gla, columbia bay
Segmental effects	misarticulation of /r/ and /l/ (0) northerly, little, drizzle, visibility /s/ becomes /sh/ see Figure 3 final devoicing (e.g. /z/ – > /s/) (-1,0,+1) Valdez – > Valdes
Suprasegmental effects	reduced speaking rate see Figures 4 & 5 mean change in pitch range (talker dependent) see Figure 6 increased F0 jitter see Figure 6

the speaker corrected himself. As indicated in the table, the only examples of gross speech effects which we found in the NTSB tapes occurred in the recording made one hour before the accident.

Segmental Phenomena

Also in Table 3, we have listed some examples of segmental errors. The problem with these data is that the recordings are noisy. Identifying most of the examples listed in the table required repeated listening and phonetic transcription (the exception is the /s/ – > /sh/ example). The amount of noise on the tape increases the probability that the transcriptions are inaccurate. Therefore, we performed acoustic analyses of several productions of /s/.

Figure 1 shows power spectra of /s/ and /sh/ produced by the first author (KJ). The horizontal axis in these graphs shows frequency from 0 to 5000 Hz and the vertical axis shows amplitude in decibels. Many speech sounds (including /s/ and /sh/) can be distinguished by their amplitude spectra because they have differing amounts of energy at different frequencies. In particular, /s/ is characterized by a peak of energy in the range from 4000 to 5000 Hz, while /sh/ has a lower frequency peak (in the range from 3000 to 4000 Hz) and a lower amplitude peak of energy in the range from 2000 to 3000 Hz. The spectra in Figure 1 illustrate what the power spectra of /s/ and /sh/ look like in recordings which have a high signal-to-noise ratio and frequency information up to 5000 Hz (see also Borden and Harris, 1984, p. 189).

Insert Figure 1 about here

Figure 2 shows power spectra of the /sh/s of *shout* and *she's* (and spectra of background noise near the fricative) as spoken by Captain Hazelwood in the recording made 33 hours before the accident. The spectra in Figure 2 give an indication of what this speaker's /sh/ will look like in this type of display. The lower amplitude peak between 2000 to 3000 Hz, illustrated in Figure 1, is present in the spectra in Figure 2, but the higher frequency information which would serve as the most reliable information distinguishing /s/ and /sh/ is not present in these spectra because the radio transmission equipment was band limited at 3000 Hz[7]. In making these comparisons, we had to be concerned also about the spectral shape of the background noise in the NTSB recordings. The spectra in Figure 1 were calculated from recordings made in a quiet recording booth, while the NTSB recordings have background noise which may be confused with fricative noise. Therefore, paired with each fricative spectrum from the recordings, we also show a spectrum of nearby background noise as a baseline against which the fricative spectrum can be compared.

---

[7] Energy above 3000 Hz was attenuated at approximately 50 dB per octave with a noise floor 50 dB below maximum signal level.

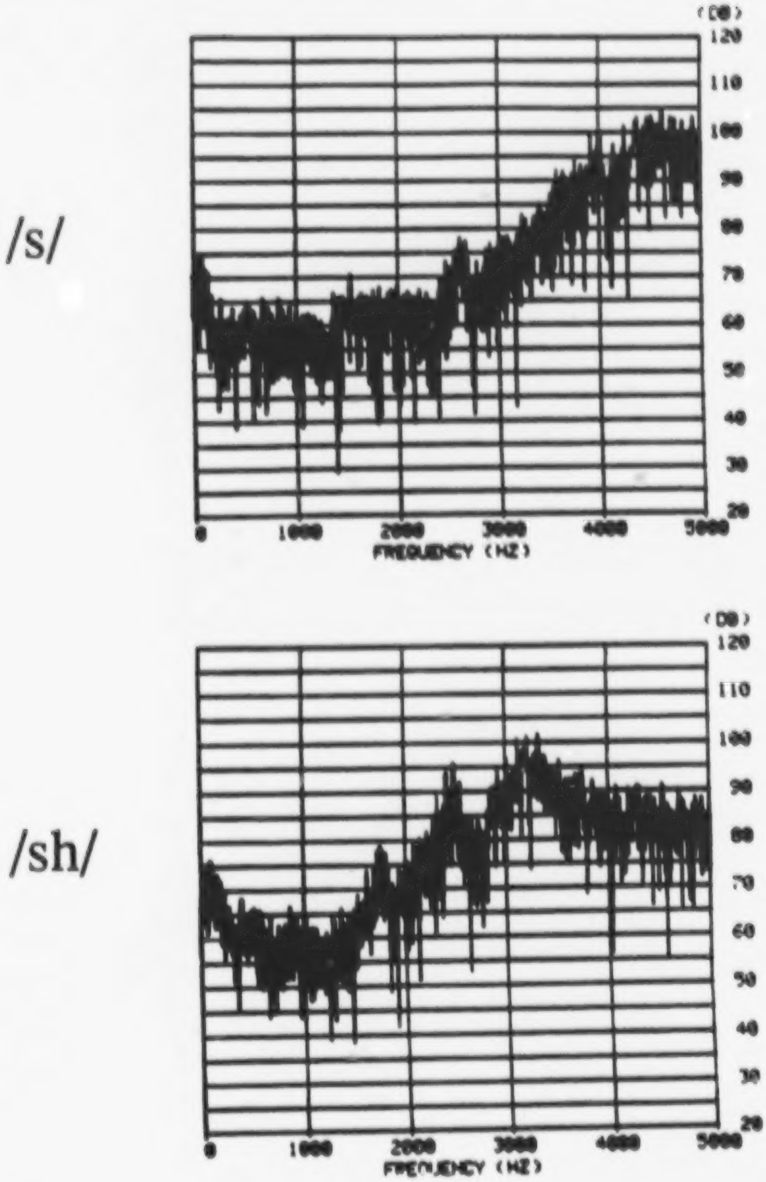


Figure 1. Power spectra of /s/ (top) and /sh/ (bottom) produced by KJ in a quiet recording booth with recording equipment responsive up to 5000 Hz.

Insert Figure 2 about here

Figure 3 shows power spectra of the /s/ of *sea* (or *see*) from the five different recordings paired with spectra of background noise from the same recording. The noise spectra were taken from nearby, open-mike background noise. On average the noise segments were 1.3 seconds from the /s/ segments[8]. The /s/ spectrum from the earliest recording (33 hours before the accident) has the same basic shape that the background noise has, suggesting that the /s/ is buried beneath the noise, or more accurately, that the main spectral energy for /s/ is not within the frequency range of the transmission system. The same is true for the /s/ of *sea* recorded one hour before the accident. The spectra of /s/ from the recordings made immediately after the accident and one hour after the accident have peaks of energy (relative to the background noise) in the region from 2000 to 3000 Hz. Finally, the spectrum of /s/ recorded 9 hours after the accident does not have a peak of energy in the region from 2000 to 3000 Hz. We interpret the peaks in the /s/ spectra from samples recorded immediately before the accident and one hour after the accident as evidence for a segmental change from /s/ to /sh/. There is no evidence in these spectra, nor in the other /s/ spectra which we examined, for this segmental change between the earliest recording and the one made one hour before the accident. These spectral changes reflect a change in the articulation of /s/ which has been observed in earlier studies of the effects of alcohol on speech production (Lester & Skousen, 1974; and Trojan & Kryspin-Exner, 1968).

Insert Figure 3 about here

Suprasegmental Properties

Finally, we examined the suprasegmental properties of the speech samples. Because the communication equipment had an automatic gain control and the distance between the microphone and the speaker's lips was (presumably) variable, it is inappropriate to compare measurements of speech amplitude or long-term average spectra. Therefore, we focussed our attention on speaking rate and voice fundamental frequency. We took care to control for discourse position and the position of words within sentences because these factors can

---

[8] We estimate that the signal-to-noise ratio in these samples ranges from 5 to 10 dB. This estimate of signal-to-noise ratio was taken from measurements of background noise during stop closures because the transmission equipment had an automatic gain control making amplitude measures from pauses inappropriate. Note also that this means that the amplitudes of the background noise spectra in Figures 2 and 3 do not accurately reflect the amplitude of background noise in the fricative spectra.

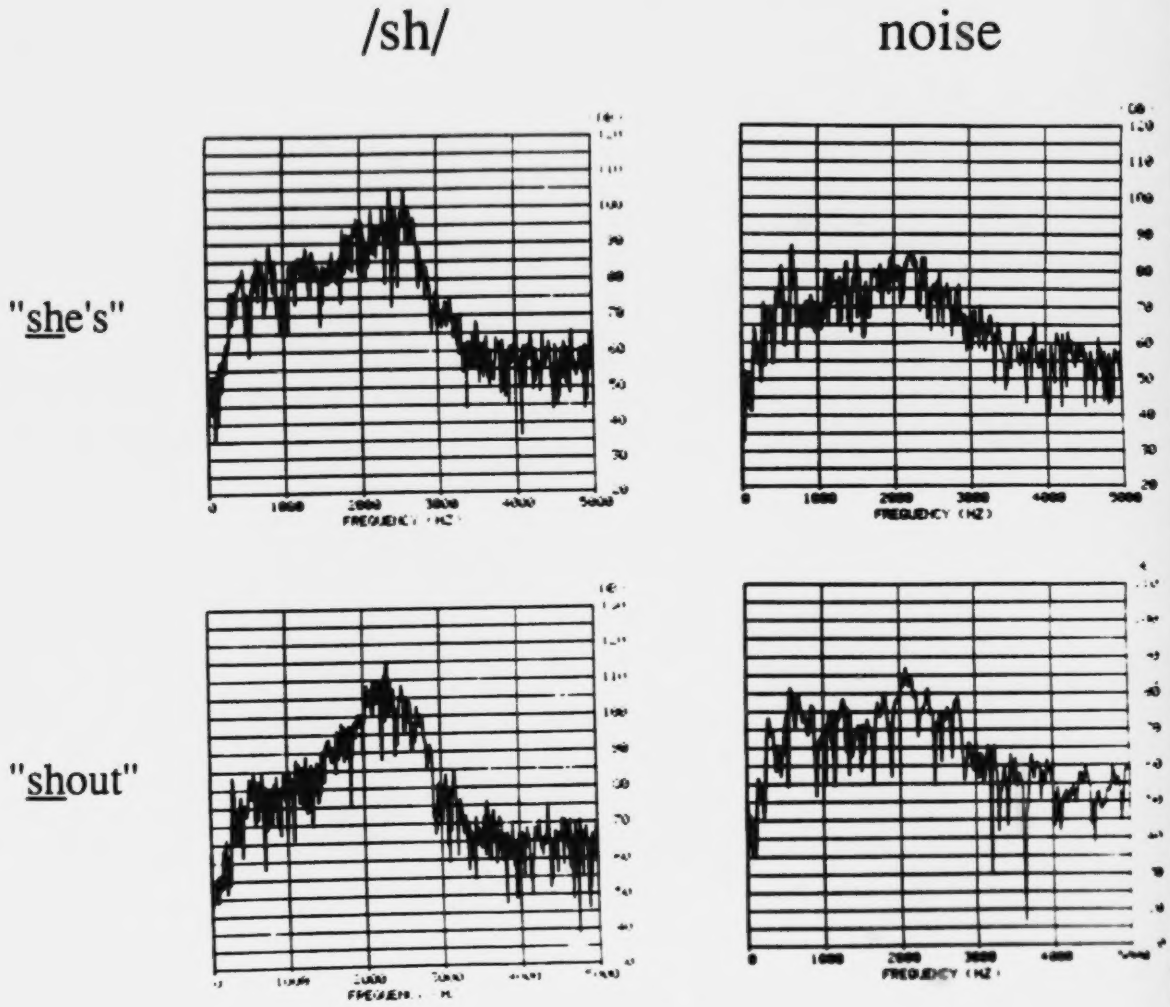


Figure 2. Power spectra of /sh/ produced by Captain Hazelwood in the words *she's* and *shout* recorded 33 hours before the accident. Each spectrum is paired with a spectrum of the background noise from a nearby open-mike pause.

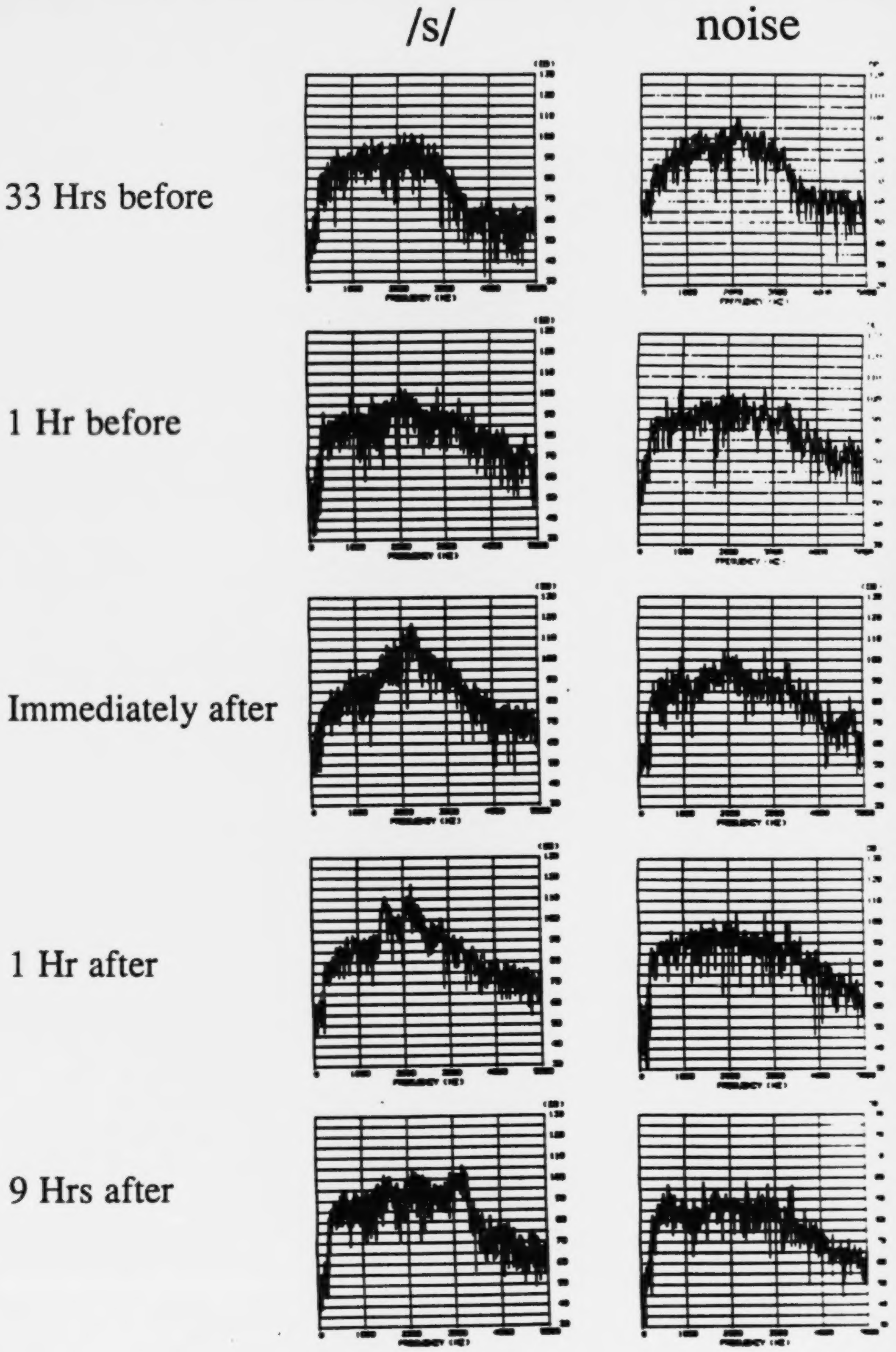


Figure 3. Power spectra of /s/ paired with spectra of nearby open-mike pauses from each of the NTSB recordings.

effect the suprasegmental properties of speech (Lehiste, 1970; Klatt, 1976). We analyzed two phrases, "*Exxon Valdez*" and "*thirteen and sixteen*", because these phrases were repeated several times during the recordings and occupied comparable positions in discourse and sentence contexts across the different recordings. Thus, these phrases provide a measure of control which is needed in making valid suprasegmental comparisons across speech samples.

Figure 4 shows durations of the speech segments in *Exxon Valdez* from each of the recordings. Each bar in this figure is the average of two occurrences of the phrase. As indicated in the top panel, it took longer to say the phrase in the samples recorded near the time of the accident. The bottom panel of Figure 4 (which is another plot of the same data) shows that this effect was more pronounced for the vowels and the /v/ of *Valdez*. If we take this as an index of speaking rate, it is reasonable to conclude from these measurements that the Captain was speaking more slowly in the samples recorded around the time of the accident than in the other samples on the NTSB tape.

Insert Figure 4 about here

One occurrence of the word *Valdez* occurred in the televised interview. This word was spoken in a discourse position which was comparable to that of *Exxon Valdez* in the NTSB recordings (utterance initial position in a short sentence). The top panel of Figure 5 compares the duration of *Valdez* in the interview with the occurrences of this word in the NTSB recordings. This comparison suggests that the Captain was speaking at his normal rate in the recording made 33 hours before the accident, and more slowly in the recordings made around the time of the accident.

We also measured the duration of the phrase *thirteen and sixteen* which occurred in discourse final position in three of the recordings (33 hours before the accident, one hour before the accident and one hour after the accident). These measurements are shown in the bottom panel of Figure 5. As with the durations of the phrase *Exxon Valdez*, this analysis indicates that Captain Hazelwood was speaking more slowly in the recordings made around the time of the accident than in the recording made 33 hours before the accident.

Insert Figure 5 about here

Durational changes are perhaps the most reliable effects we have found in the NTSB recordings and they suggest that Captain Hazelwood was speaking more slowly than normal around the time of the accident. These changes in duration are consistent with the laboratory

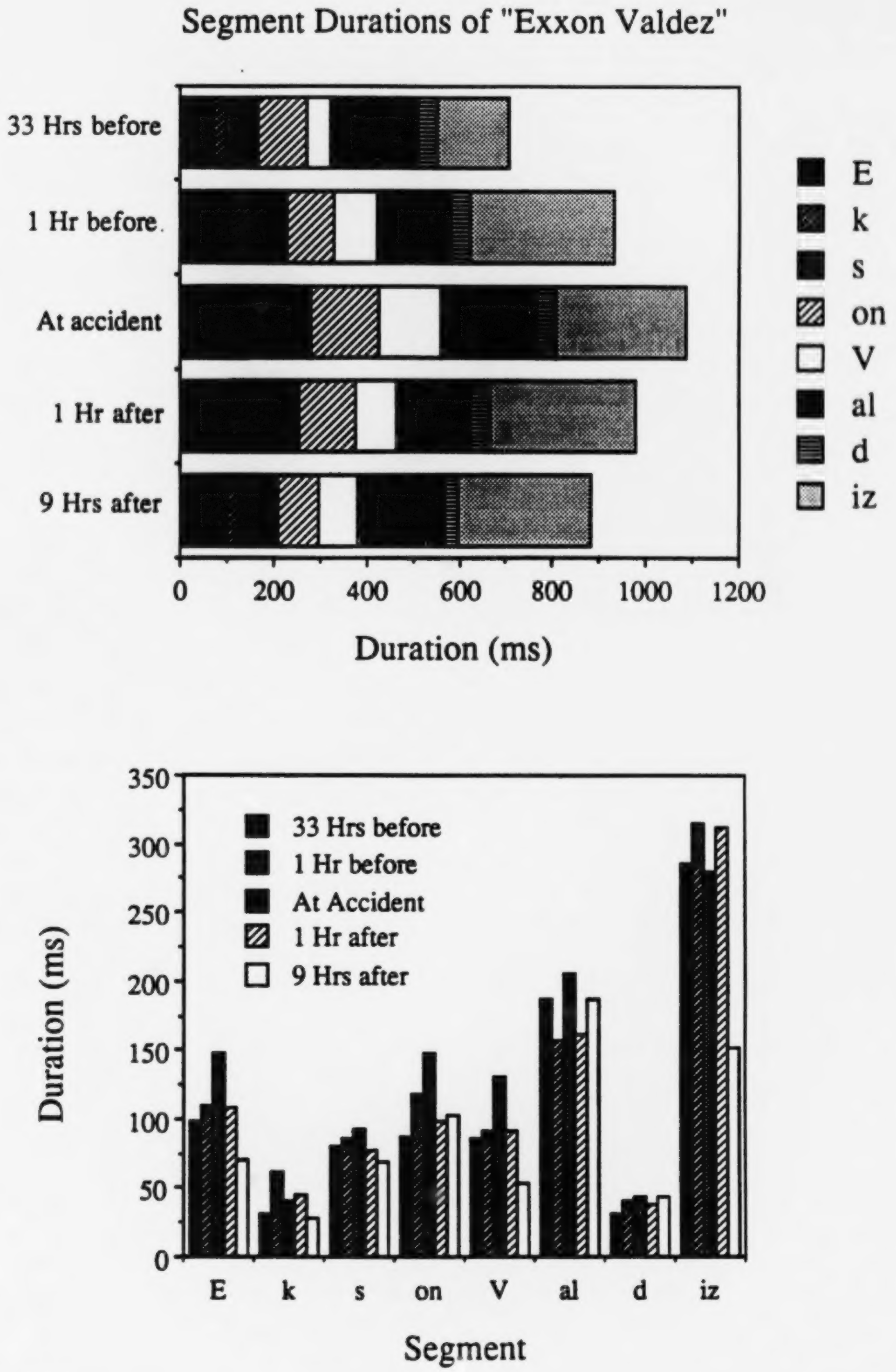


Figure 4. Durations of speech segments in the phrase, *Exxon Valdez* at the different times of recording. Top panel: cumulative durations indicating the overall increase in duration. Bottom panel: durations grouped by segments showing which segments had increased duration around the time of the accident.

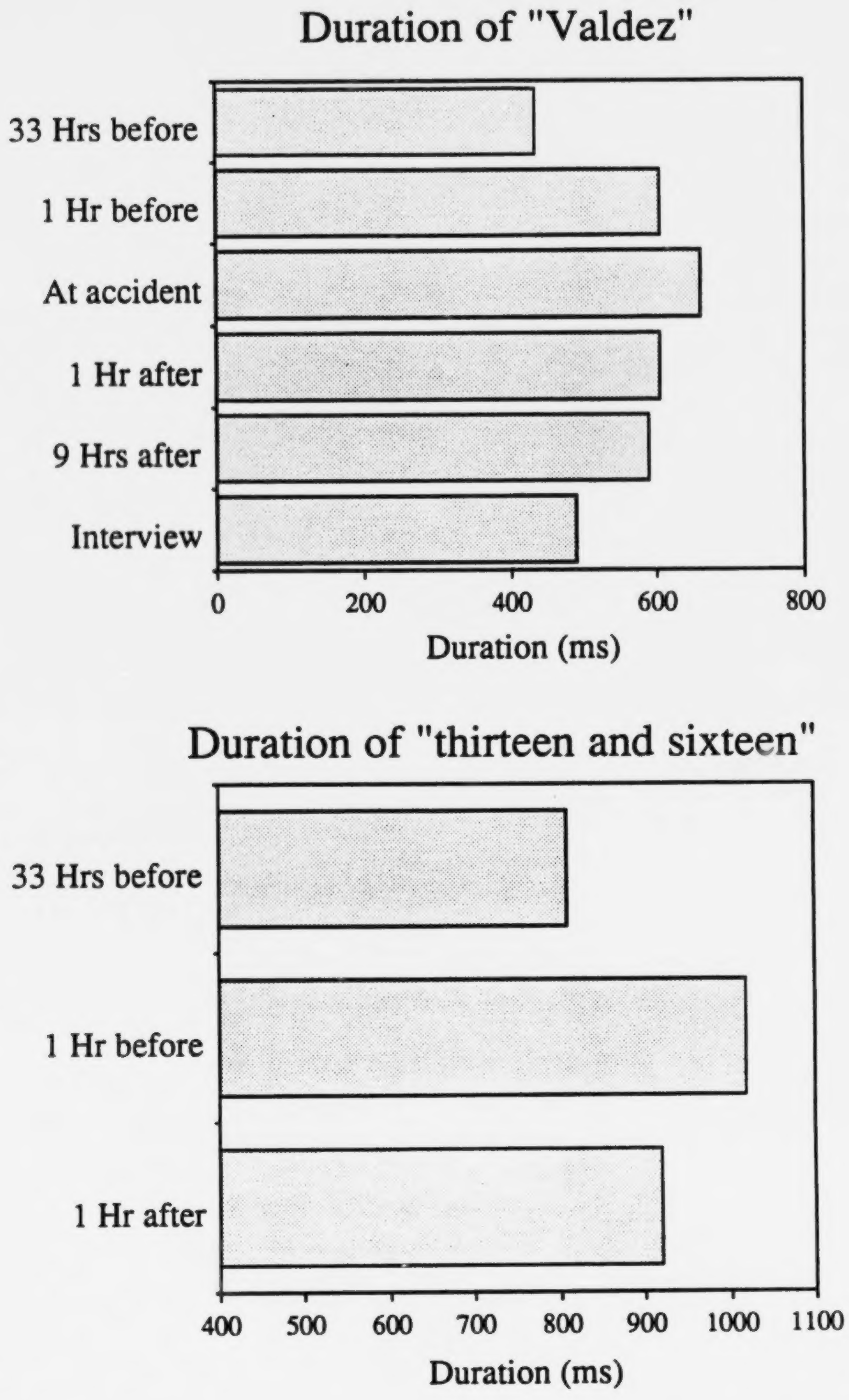


Figure 5. Top panel: Duration of the word *Valdez* from the NTSB tapes (data is the same as that in Figure 4) compared with the same word produced in a similar discourse position in the televised interview. Bottom panel: Duration of the phrase *thirteen and sixteen* from recordings made at three times around the time of the accident.

findings reported by Pisoni, et al. (1986) and Pisoni & Martin (1989) for speech produced while intoxicated.

The top panel of Figure 6 shows voice fundamental frequency (F0) averaged across the phrase *Exxon Valdez* in each of the speech samples, the phrase *thirteen and sixteen* from three of the NTSB recordings, and one sentence from the televised interview. We took F0 measurements from each of the four vowels in *Exxon Valdez* (which occurred at least twice in each of the NTSB recordings). We were not able to measure F0 in all of the vowels in *thirteen and sixteen* because this phrase occurred in utterance final position in the recordings and was produced with quite low amplitude. Each point in Figure 6 for *thirteen and sixteen* is based on measurements from at least two vowels. The last point in each panel shows data averaged across a sentence in the televised interview[9].

The normal pitch detection algorithms were unable to operate on the NTSB speech samples because of the degree of background noise; therefore, we modified an existing vocal jitter algorithm (see Pinto & Titze, 1990 for a recent review). We adapted the existing technique by rectifying and low-pass filtering the signal (to remove high frequency noise) before attempting to find successive pitch periods. The results of the algorithm were visually confirmed and then F0 and jitter measures calculated. We calculated Davis' (1976) pitch perturbation quotient (PPQ) which is the ratio of the "average perterbation measured from the pitch period" and the average pitch period (p. 51, 123).

As the top panel shows, voice fundamental frequency was dramatically lower in the samples recorded around the time of the accident.[10] Also, this panel shows the average F0 range in each speech sample. The different samples cannot be distinguished by their F0 range (except perhaps the items from the recording made nine hours after the accident), but there was a trend for items near the time of the accident to have more F0 jitter (bottom panel of Figure 6). This finding is consistent with Pisoni & Martin's (1989) observation that speakers had higher standard deviation of F0 after alcohol consumption. (Note the discussion above concerning the ways in which SD F0 may be affected.) The lower jitter in the sentence taken from the televised interview (CC) is consistent with Brenner et al.'s (1985) observation that talkers have less F0 jitter when in stressful situations.

Insert Figure 6 about here

In summary, the acoustic-phonetic measurements presented here are all consistent with the findings of previous controlled laboratory studies of the effects of alcohol on speech pro-

---

[9] The sentence was, "I would say the same for the state of Alaska, they came after me, hammer and tong."

[10] Fundamental frequency as low as that seen here normally occurs only in a mode of vocal cord vibration called creak, or pulse register phonation. In English this mode of vocal cord vibration is usually seen only at the ends of declarative sentences, although this varies somewhat from speaker to speaker.

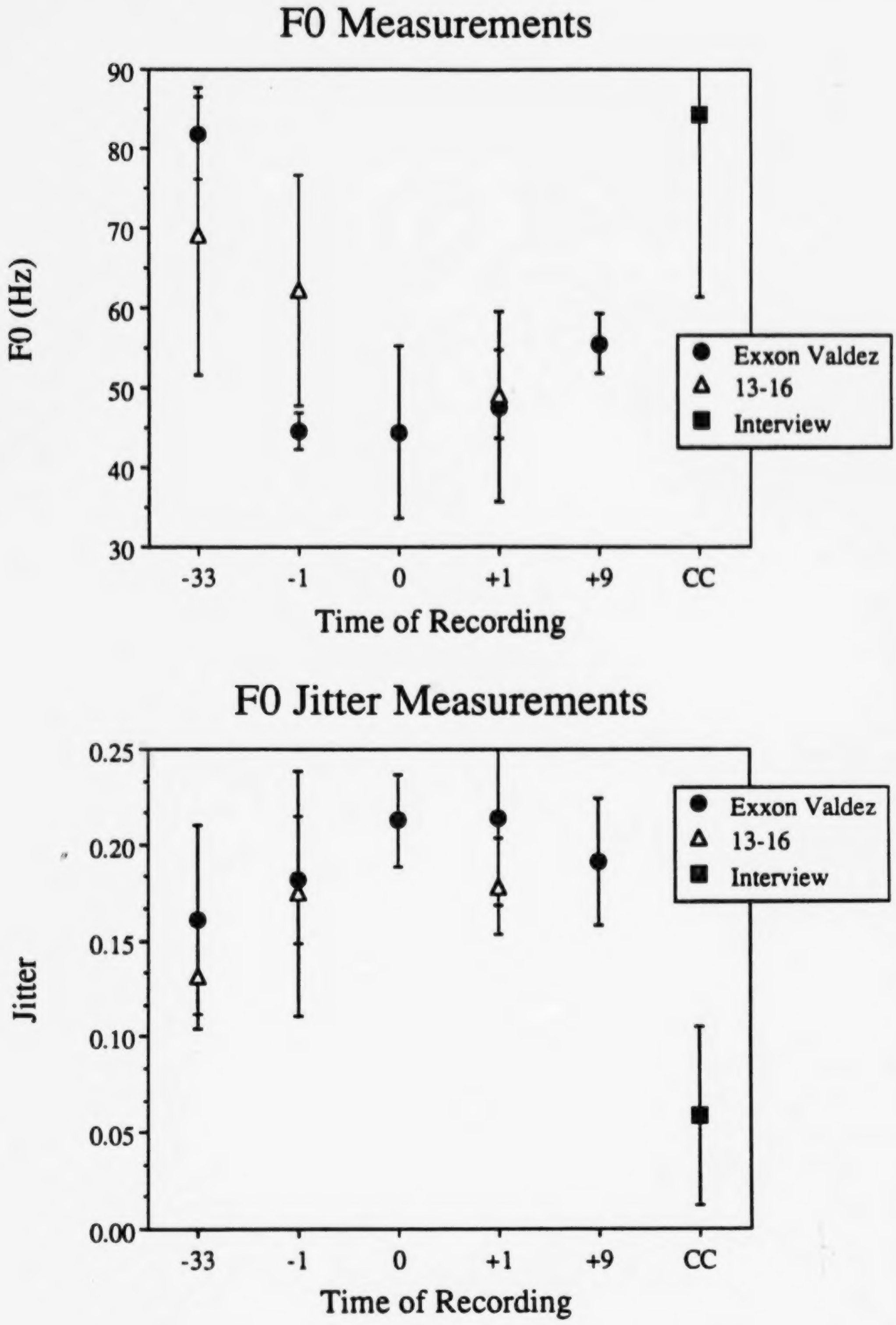


Figure 6. Top panel: Average fundamental frequency (pitch of voice) in *Exxon Valdez, thirteen and sixteen* and from one sentence in the televised interview as a function of time of recording. Bottom panel: F0 jitter measurements from the same speech samples.

duction. In listening to the recordings, we observed a number of gross misarticulations and segmental misarticulations around the time of the accident. We also found acoustic evidence in two of the recordings made near the time of the accident (0, +1) for a misarticulation of /s/. Finally, we found that Captain Hazelwood was speaking more slowly, and used a lower fundamental frequency with more fundmental frequency jitter around the time of the accident as compared with his speech 33 hours before the accident and his speech in the televised interview.

Conclusions

We now return to the theme with which this report began. Is it possible to determine, from acoustic analyses of speech, whether an individual is intoxicated? We have presented *a priori* arguments that it is. We also found in a review of previous research on environmental and emotional effects on speech production, that the effects of alcohol are unique among the previous findings. In our present analyses, we have also found a pattern of changes in Captain Hazelwood's speech which is consistent with the pattern of changes observed in previous laboratory studies on the effects of alcohol on speech production (this was as much as we concluded in our preliminary report). Taken together, these findings suggest that the Captain was intoxicated at the time of the accident. There are, however, several methodological and empirical problems that must be taken into consideration with regard to this conclusion.

First, there are gaps in the previous research; both in research concerning the effects of alcohol on speech production and in research on other effects on speech production. For instance, we have reported here measurements of vocal jitter. This is the first time that vocal jitter measurements have been reported in the context of a study of the effects of alcohol on speech. We also noted several gaps in previous research on environmental and emotional effects on speech. For instance, we are not aware of any research which has attempted to explore the effects of fatigue on speech, or any research which explores the ways in which various environmental and/or emotional factors may interact in their effects on speech. In the absence of these types of additional data, we cannot rule out a number of other possible causes for the changes we have observed in Captain Hazelwood's speech.

Second, in addition to a lack of breadth in the existing knowledge, there is a lack of depth. There are no normative data on the effects of alcohol on speech production. We don't know how general the effects summarized in Table 1 are. Normative data are also unavailable for the effects summarized in Table 2. This lack of data makes it impossible to make reliable probabilistic statements such as, "Captain Hazelwood had this pattern of changes and 95% of the people who exhibited this pattern were intoxicated while only 10% of fatigued speakers show this pattern." Currently, statements of this type are based on studies which employed very small numbers of talkers.

Third, the recordings which we were working with in the present case limited the type and quality of the measurements we could make. For instance, it would have been very informative to know whether the Captain was speaking more loudly or softly in the recordings near the time of the accident. This measure was not possible with the NTSB recordings because automatic gain control was used in the transmission equipment and the placement of the microphone in relation to the speaker's lips was (presumably) variable. Furthermore, the variability of the background noise made the calculation of long-term average (LTA) spectra invalid, though Klingholz et al. (1988) found reliable changes in LTA spectra when speakers were intoxicated. Our analysis of fricative spectra was also hampered by the presence of background noise and the frequency response characteristics of the transmission equipment. Finally, the complicated history of the recording made 33 hours before the accident casts some doubt on the measurements taken from that recording. We have outlined the magnitude of error which may have resulted from this situation and have taken measurements from a televised interview to serve as another "control" condition. Still, this extra link in the history of the recording introduces an additional source of error that would not have existed if the original Coast Guard recording had not been erased.

A number of aspects of the data we have reported here suggest that Captain Hazelwood was intoxicated when the Valdez ran aground. Especially suggestive is the pattern that we have observed in measurements of four different speech parameters. The changes in F0, F0 jitter, duration and fricative spectra measurements are all consistent with the hypothesis that Captain Hazelwood was intoxicated at the time of the accident. These four parameters also have an inflection point around the time of the accident. This, coupled with the knowledge that the Captain's blood alcohol level ten hours after the accident was 0.06%, suggests that his blood alcohol level may have been higher at the time of the accident. In addition to these fine-grained acoustic analyses, we also found some additional segmental misarticulations and some gross errors in the recordings made around the time of the accident. From these findings, we conclude that Captain Hazelwood displayed changes in sensory-motor behavior that are similar to those found in earlier laboratory based studies in which the talkers were intoxicated to known BALs. This similarity suggests that the Captain was intoxicated at the time of the accident. However, this conclusion should be qualified in light of the limitations of the present recordings and the limited scientific data on the effects of alcohol and other variables on speech production.

References

Anttila, R. (1972). *An introduction to historical and comparative linguistics.* MacMillan, New York.

Berry, M.S. & Pentreath, V.W. (1980). The neurophysiology of alcohol. In Sandler, M. (Ed.) *Psychopharmacology of alcohol* (pp. 43–72). Raven Press, New York.

Brenner, M. & Shipp, T. (1988). Voice stress analysis. In *Mental-state estimation 1987* (pp. 363–376). NASA Conference Publication 2504.

Brenner, M., Shipp, T., Doherty, E.T. & Morrissey, P. (1985). Voice measures of psychological stress: Laboratory and field data. In Titze, I.R. & Scherer, R.C. (Eds.) *Vocal fold physiology, biomechanics, acoustics, and phonatory control* (pp. 239–248). The Denver Center for the Performing Arts, Denver.

Davis, S.B. (1976). Computer evaluation of laryngeal pathology based on inverse filtering of speech. *SCRL Monograph*, **13**.

de Villiers, J.G. & de Villiers, P.A. (1978). *Language acquisition.* Harvard University Press, Cambridge, MA.

Dunker, E. & Schlosshauer, B. (1964). Irregularities of the laryngeal vibratory pattern in healthy and hoarse persons. In Brewer, D.W. (Ed.) *Research potentials in voice physiology* (pp. 151–184). International Conference at Syracuse, 1961, Syracuse, N.Y.

Griffin, G.R. & Williams, C.E. (1987). The effects of different levels of task complexity on three vocal measures. *Aviation, space, and environmental medicine*, **58**, 1165–1170.

Hansen, J.H.L. (1988) *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition.* PhD Dissertation, Geogria Institute of Technology.

Hellekant, G. (1965). The effect of ethyl alcohol on non-gustatory receptors of the tongue of the cat. *Acta Physiology Scandinavia*, **65**, 243–250.

Jakobson, R. (1941). *Kindersprache, Aphasie und Allgemeine Loutgesetze.* Uppsala.

Klatt, D.H. (1976). Linguistic uses of segmental duration in American English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208–1221.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *Journal of the Acoustical Society of America*, **84**, 929–935.

Lehiste, I. (1970). *Suprasegmentals.* MIT Press, Cambridge, MA.

Lester, L. & Skousen, R. (1974). The phonology of drunkenness. In Bruck, A., Fox, R., & LaGaly, M. (Eds.) *Papers for the parasession on natural phonology* (pp. 233-239). Chicago Linguistic Society.

Lindblom, B. (1983). Economy of speech gestures. In MacNeilage, P.F. (Ed.) *The production of speech* (pp. 217-245). Springer-Verlag, New York.

Lisker, L. & Abramson, A.D. (1964) A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, **20**, 384-422.

Moore, T.J. & Bond, Z.S. (1987). Acoustic-phonetic changes in speech due to environmental stressors: Implications for speech recognition in the cockpit. Presented at the *4th Annual symposium on aviation psychology* (pp. 26-30), April, 1987.

Pinto, N.B. & Titze, I.R. (1990). Unification of perturbation measures in speech signals. *Journal of the Acoustical Society of America*, **87**, 1278-1289.

Pisoni, D.B., Bernacki, R.H., Nusbaum, H.C. & Yuchtman, M. (1985) Some acoustic-phonetic correlates of speech produced in noise. *Proceedings of IEEE ICASSP*, pp. 1581-1584.

Pisoni, D.B., Hathaway, S.N., & Yuchtman, M. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. *Alcohol, accidents and injuries, (Society of Automotive Engineers, Pittsburgh, PA), Special Paper P-173* (pp. 131-150).

Pisoni, D.B. & Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, **13**, 577-587.

Sobell, L. & Sobell, M. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, **15**, 861-868.

Sobell, L., Sobell, M., & Coleman, R. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica*, **34**, 316-323.

Stevens, K.N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In David, E.E. & Denes, P.B. (Eds.) *Human communication: A unified view*. McGraw-Hill, New York.

Streeter,L.A., MacDonald, N.H., Apple, W., Krauss, R.M. & Galotti, K.M. (1983) Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, **73**, 1354-1360.

Subtelny, J.D., Oya, N. & Subtelny, J.D. (1972). Cineradiographic study of sibilants. *Folia Phoniatrica*, **24**, 30-50.

Trojan, F. & Kryspin-Exner, K. (1968). The decay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica*, **20**, 217–238.

Wallgren, H. & Barry, H. (1970). *Actions of alcohol.* Vol. I. Elsevier, Amsterdam.

Williams, C.E. & Stevens, K.N. (1972) Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, **52**, 1238–1250.

Williams, C.E. & Stevens, K.N. (1981). Vocal correlates of emotional states. In Darby, J.K. (Ed.) *Speech evaluation in psychiatry* (pp. 221–240). Grune & Stratton, New York.

Effects of Talker Familiarity on Serial Recall of Spoken Word Lists[1]

Nancy Lightfoot

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] **I thank David Pisoni, Steve Goldinger and Keith Johnson for their comments on an earlier version of this paper. This research was supported, in part, by NIDCD Research Grant R01 DC0111-13 to Indiana University in Bloomington, IN.**

Abstract

Compared to lists of words produced by a single talker, lists produced by multiple talkers lead to decreased accuracy of serial recall in the primacy portion of the recall curve at fast presentation rates. This pattern of recall for multiple and single talkers reverses at slow presentation rates (Goldinger, Pisoni & Logan, under review). Explanations for these differences in recall have proposed that processing multiple-talker stimuli either increases rehearsal demands during transfer to long-term memory (LTM) or increases difficulty of perceptual encoding. The present experiments examine these competing explanations for differential serial recall of multiple- versus single-talker word lists using voices that subjects have been trained to identify. As listeners become familiar with a talker's voice, it is possible either that mandatory rehearsal will encompass richer recall cues, including talker-specific information, or that perceptual normalization will become more automatic. Results provide support for the proposal that rehearsal of enriched cues leads to the differences in recall of multiple- and single-talker lists. Using familiar voices increases serial recall for multiple-talker lists over single-talker lists in the primacy portion of the recall curve, but does not appear to facilitate perceptual encoding. Primacy recall for multiple-talker lists is further enhanced at a slower presentation rate.

Effects of Talker Familiarity on Serial Recall of Spoken Word Lists

The physical representation of the speech signal varies widely depending on phonemic and sentential context (Pisoni & Luce, 1987; Mullennix, Pisoni, and Martin, 1989), and on the dialect, age, gender, and vocal tract length of the speaker (Mattingly, Studdert-Kennedy, & Magen, 1983; Mullennix, et al., 1989). The process of talker normalization involves extracting a consistent linguistic representation of speech from this highly variable speech signal. Intuitively, it seems that talker normalization is automatic and cost-free. Recent research, however, indicates that increasing talker variability leads to measurable deficits in information processing. Same-different judgement tasks (Mullennix & Pisoni, 1988), word identification (Mullennix, et al., 1989), and serial recall from the primacy portion of the serial position curve (Martin, Mullennix, Pisoni & Summers, 1989; Logan & Pisoni, 1987), all seem to be more difficult for stimuli produced by multiple as opposed to single talkers. In an interesting new study, however, Goldinger, Pisoni, & Logan (under review) have shown advantages for recall for multiple-talker stimuli by slowing the presentation rate of to-be-remembered (TBR) word lists. The present research seeks to replicate the findings of Goldinger et al., and to offer additional insight into the role of talker-specific information in serial recall.

The classic distinction between short and long-term memory stores has been supported, in part, by the finding that numerous manipulations differentially affect the primacy and recency portions of the serial position curve (Glanzer & Cunitz, 1966; Sumby, 1963). There is some evidence that the explanation for increased primacy recall lies in an increased number of rehearsals for early list items (Greene, 1986; Rundus, 1971), leading to increased probability of transfer to long-term memory. In support for this theory of primacy recall, Rundus (1971) used an overt rehearsal task to demonstrate that the number of rote rehearsals an item receives is positively correlated with the probability of recall across the primacy region of the serial position curve (see Shimizu, 1987, and Modigliani & Hedges, 1987 for alternative views). The primacy effect has also been shown to disappear when rehearsal is prevented (Glenberg, Bradley, Stevenson, Kraus, Tkachuk, Gretz, Fish, & Turpin, 1980).

Another variable which has been shown to selectively affect primacy recall, however, is the depth of stimulus processing, or the degree of elaborative rehearsal that each stimulus receives (Craik & Lockhart, 1972): the richer the processing of each stimulus, the better the recall for that stimulus, at least in initial positions of a word list. Shimizu (1987), for example, has shown that while the number of rote rehearsals in a overt rehearsal paradigm increases with slowed presentation rates, this increased rehearsal was not correlated with long-term recall. The nature of the rehearsal process and distinctiveness of encoding for TBR material might provide a better index of primacy recall. In fact, primacy effects in recall probably reflect an interaction of rote rehearsal and depth of processing (Crowder, 1976). Atkinson and Shiffrin (1968), for example, propose that rote rehearsal minimally maintains information in short term memory while other items are being processed elaboratively and transferred to long-term memory. Maintenance rehearsal does seem to play an important role in recall from short-term memory: recency effects are thought to represent the "dumping out" of the contents of short term memory, which are maintained through rote rehearsal (see Bjork & Whitten, 1974 for an alternative view).

With regard to recall of spoken word lists, the process of talker normalization could operate on either the efficiency of maintenance rehearsal, or elaborative rehearsal processes, or both in a memory task. Previous studies have shown that the process of rehearsal and transfer to long-term memory is different for multiple- and single-talker stimuli. One possible explanation for these differences is that multiple-talker stimuli are simply more difficult to encode, which leads to a time lag before rehearsal of each item can begin. This time-lag could affect the efficiency of maintenance rehearsal for multiple-talker word lists, leading to a greater loss of information in processing of multiple-talker stimuli. An encoding delay could interact with the depth of elaborative rehearsal as well. An alternative account of multiple- and single-talker recall differences, proposed by Goldinger et al. (under review), is that talker-specific information provides a richer set of cues for elaborative processing. Talker information may, in fact, demand mandatory processing and rehearsal (Mullennix & Pisoni, 1988; Mullennix & Pisoni, in press), leading to enhanced rehearsal demands for lists composed of multiple-talker stimuli.

There is some support, however, for the hypothesis that multiple-talker stimuli are more difficult to perceive or encode than comparable single-talker stimuli. Research in the area of talker normalization has shown that processing of words and vowels produced by multiple as opposed to single talkers does exert observable costs in perceptual tasks (Verbrugge, Strange, Shankweiler, & Edman, 1976; Cole, Coleheart, & Allard, 1974; Creelman, 1957, Mullennix & Pisoni, in press). Mullennix et al. (1989), for example, investigated the process of talker normalization in word recognition and attempted to isolate the level of processing at which talker normalization occurs in speech perception. Mullennix et al. manipulated variables thought to affect low level accoustic processing (signal degradation and signal-to-noise ratio), and variables thought to interact with lexical access (word frequency and lexical density) for words produced by either multiple talkers or a single male talker. They found that the multiple- versus single-talker condition interacted reliably with low level perceptual interference, but not with changes in word frequency or lexical density. Mullennix et al. concluded that talker normalization may be independent of processes used in word recognition or lexical access, although this conclusion was based on a failure to reject the null hypothesis.

There is also evidence that perceptual costs in speech processing may divert limited capacity resources from higher level cognitive processes: the perceptual demands associated with talker normalization may "cascade up" the speech processing system (Luce, Feustel, & Pisoni, 1983; Mattingly, Studdert-Kennedy, & Magen, 1983; Craik & Kirsner, 1974; Martin et al., 1989; Logan & Pisoni, 1987). These processing costs may then affect encoding in memory or efficiency of rehearsal, presumably impairing transfer of to-be-remembered items to long-term memory. This account is supported by the finding that synthetic-speech word lists, as well as multiple-talker word lists, produce pronounced deficits in serial recall in the primacy region of the serial position curve (Luce et al., 1983; Mattingly et al., 1983; Martin et al.,1989; Logan & Pisoni, 1987). Luce et al. (1983), for example, examined serial recall of lists of synthetic and natural speech and found reliable decrements in the primacy portion of the learning curve for recall of synthetic speech. While synthetic speech is probably more difficult to perceive than natural speech (e.g. Luce et al., 1983), there is little reason to believe that consistent synthetic speech characteristics would lead to better serial recall cues than natural speech produced by a single talker. Presumably multiple-talker information acts by increasing the distinctiveness of memory cues, or by providing distinctive order cues for serial recall. Because synthetic speech is uniform within lists, however, it is unlikely to provide unique recall cues for word or order information. This explanation remains to be tested however.

Martin and his colleagues (Martin et al., 1989) examined the effects of talker variability on memory for lists of isolated words. Like Luce et al. (1983), Martin et al. found that the process of perceptual normalization produced decrements in the primacy portion of the serial position curve for a serial recall task. Using a free recall paradigm, however, Martin et al. found no significant effects for talker variability. Martin et al. explained these results by suggesting that serial recall imposed additional processing demands on subjects compared with a free recall task. In serial recall subjects must encode both the item and order information. This additional cognitive load presumably combined with processing costs for talker normalization in the serial recall task, leading to decreased efficiency of rehearsal and decreased primacy recall. In an additional experiment, Martin found impaired recall for preload digits as a function of talker variability, again indicating increased processing demands for talker normalization in multiple-talker lists.

Martin et al. (1989) proposed two alternative explanations for decreased primacy effects in subjects who heard multiple-talker word lists. They suggested this effect could be explained either by impaired efficiency of rehearsal and transfer to long-term memory, or by less effective retrieval of acoustic cues from residual short term memory in the multiple-talker condition. In a final experiment, Martin et al. attempted to eliminate the contribution of talker specific acoustic cues from short term memory by introducing a distractor period between list presentation and recall. The length of the distractor task had no effect on the primacy portion of the curve in either the multiple-talker or single-talker conditions, although the recency effect was selectively decreased as the length of the distractor period increased. Based on these results, Martin et al. rejected an explanation based on short term memory acoustic cues, and attributed the decreased primacy effect for the multiple-talker condition to impaired efficiency of rehearsal and transfer to long-term memory.

Several other recent studies have further elaborated the differences in perception and memory of multiple- and single-talker stimuli. Mullennix & Pisoni (1988, in press) used multiple-talker stimuli in a Garner speeded-classification paradigm. Their results indicate that phonetic information and talker-specific voice cues may be processed as integral dimensions: subjects found it very difficult to selectively ignore voice information and process only phonetic cues, even when the task required them to attend to phonetic differences alone. Mullennix & Pisoni (1988) investigated the duration of multiple-talker interference in a same-different judgement task over interstimulus intervals (ISIs) of 100 to 4000 milliseconds. They found that talker information was retained and processed mandatorily for periods of up to four seconds. Cole et al. (1974) found similar results for periods of up to eight seconds. Geiselman & Bellezza (1976) demonstrated long-term recall for voice information in an incidental recall task. All of these results suggest automatic processing, rehearsal, and recall of talker-specific cues. The results of these studies are also compatible with the theory that multiple-talker voice cues demand mandatory attentional processing. It is possible that differences in primacy recall for multiple-talker stimuli reflect additional rehearsal demands for richer auditory cues, instead of, or in addition to, additional perceptual encoding demands for multiple-talker stimuli.

Goldinger et al. (under review) specifically examined the hypothesis that multiple-talker stimuli require increased rehearsal demands. Goldinger et al. manipulated two types of factors in a serial recall experiment using multiple- and single-talker word lists: factors affecting rehearsal efficiency and factors affecting ease of encoding. Goldinger et al. argued that if increased rehearsal demands due to richer cues lead to deficits for multiple talker lists, the differences between recall for multiple and single talkers should interact strongly with manipulations of rehearsal time. Specifically, multiple-talker lists should show greatly depressed recall in the primacy portion of the curve at fast presentation rates, where encoding talker information will overload the system and interfere with rehearsal of the words themselves. At slow presentation rates, on the other hand, this effect should disappear: with longer rehearsal intervals, both word and talker information may be transferred to long-term memory, resulting in more elaboration, thus providing an extra set of recall cues for item or order information.

Goldinger et al. (under review) argued that if multiple-talker deficits were due only to increased encoding demands, on the other hand, the multiple-talker condition would still lead to larger deficits in the primacy region of the curve at fast rates (see Sumby, 1963). But talker condition should also interact with other manipulations of encoding difficulty. To test this hypothesis, Goldinger et al. used two types of lists in their serial recall task: lists composed of easy-to-recognize words and lists composed of hard-to-recognize words. Words that were easy to recognize were defined as high frequency words from sparse lexical neighborhoods composed mostly of low frequency words. Hard-to-recognize words were low frequency words from high-density, high frequency neighborhoods. Under the hypothesis that multiple-talker recall will be depressed because of difficulties in perceptual encoding, the talker manipulation was expected to interact strongly with this confusibility manipulation. Specifically, the difference between multiple and single talkers should be much more marked for "hard" word lists than for "easy" word lists.

Goldinger et al.'s results unambiguously favored a rehearsal-based explanation for multiple- versus single-talker differences in primacy recall. The pattern of differences in recall for multiple- and single-talker lists changed dramatically with presentation rate. At fast presentation rates, single-talker word lists were recalled much more accurately in the primacy portion of the recall curve. At slower presentation rates, however, this pattern showed an unexpected reversal: multiple-talker word lists showed a marked advantage in early list positions, presumably due to more elaborative rehearsal, or to the additional recall and temporal order cues provided by the variations in talker information within lists. Goldinger et al. found no interaction, on the other hand, between the talker manipulation and the confusibility manipulation, or between the rate manipulation and the confusibility manipulation.

The present research provides a further test of Goldinger et al.'s (under review) hypothesis. In these studies, the availability of talker specific cues for serial recall was manipulated in two ways. In the first experiment, familiarity with the talkers in the recall lists was manipulated between subjects. Training subjects to exploit talker specific cues in an identification task should show some degree of transfer to a serial recall task using word and voice information. In the second experiment, serial recall for familiar voices was examined at a slow presentation rate. If slowing presentation allows more complete rehearsal and transfer of items to long-term memory, as Goldinger et al. (under review) have suggested, there should be even greater advantages for serial recall for multiple-talker lists at this slower rate compared to single-talker lists.

In Experiment 1, half of the subjects were familiarized with the voices of the single and multiple talkers over a period of two weeks of identification training. Subjects in the trained condition demonstrated a stable, long-term representation of these talkers' voices by achieving a consistent, high degree of accuracy in identifying the talkers. Following identification training, the trained subjects and a group of naive subjects were run in a serial recall experiment. The presentation rate used in Experiment 1 was compatible with earlier studies showing reliable multiple-talker decrements in primacy recall. For the trained subjects, talker-specific recall cues should be much more accessible, even at faster presentation rates, leading to increased multiple-talker performance in the primacy portion of the serial recall curve. Alternatively, familiarity might have no effect, or might have its effect at a low-level of perceptual encoding. Familiarity with voices, for example, might allow subjects to recognize words produced by different talkers more easily, because of familiarity with the idiosyncracies of the particular speech patterns of those talkers. In this case, an interaction with a confusibility manipulation might be expected. As in Goldinger, et al's (under review) study, the present experiment included a confusibility manipulation as a within subjects variable. This confusibility manipulation was included to test for interactions between the talker manipulation and factors associated with encoding difficulty. If confusibility interacts with the talker condition or with the training manipulation, this would provide evidence that perceptual or encoding difficulties are responsible for differences in recall for multiple-

Experiment 1

Method

Subjects. Subjects were 36 paid volunteers from the Bloomington, Indiana community. All participants were native speakers of English who reported no history of a speech or hearing disorder at the time of testing. Subjects were tested individually or in groups of two to six in testing booths in a sound-treated room. Trained subjects participated in a two-week training program prior to the memory test. Untrained subjects participated in the memory test only.

Materials. Two sets of stimuli were used in the course of the experiment: Memory stimuli and training stimuli. Memory stimuli included a subset of words from the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). These stimuli were selected, recorded, and digitized by Logan and Pisoni (1987). All stimuli were monosyllabic words that received average ratings of "highly familiar" in a previous study by Nusbaum, Pisoni, and Davis (1984). In the multiple-talker condition, stimuli were recorded by ten different talkers, five of whom were female and five of whom were male. Words in the single-talker condition were produced by a single male speaker who was also represented in the multiple talker lists. Training stimuli were 150 phonetically balanced, single-syllable words produced by the same ten talkers who had produced the memory stimuli. All speech stimuli were digitized on a PDP 11/34 computer using a 12-bit analog-to-digital converter. RMS amplitude across words was equated using a signal processing package. Stimuli were presented to subjects over TDH-39 headphones at a level of 80 dB SPL (a comfortable listening level). Digitized stimuli were reproduced using a 12-bit digital-to-analog converter and were low pass filtered at 4.8 kHz before presentation. A PDP 11/34 computer was used to generate word lists and to control the presentation of stimuli in real-time (Logan & Pisoni, 1987).

Procedure

Training. All trained subjects completed nine training sessions over a period of two weeks prior to the final memory test. A training session, which lasted approximately one hour, consisted of two learning cycles and a test cycle. During each learning cycle, subjects were first exposed to the voices (familiarization phase) and then attempted to identify the voices with trial-by-trial feedback on their performance (training phase). During the familiarization phase, subjects heard five word lists from each of the ten talkers in succession. They then heard the same lists a second time. Finally they heard a ten-word list composed of one word from each talker in succession. Each time a token was presented to subjects during familiarization, the name of the appropriate talker was displayed on the center of the CRT screen in front of them. This familiarization procedure was intended to help subjects develop a robust representation of the talkers' voices by giving them some direct experience with the degree of variability within each talker's speech, and by juxtaposing the talkers' voices against each other.

During the training phase of the learning cycles, subjects practiced identifying the talkers who had produced each token. The number keys on the CRT response board were labelled with ten common English names. The keys 1-5 were labelled with male names (John, Mike, Tom, Brad, Dan), and the keys 6-10 were labelled with female names (Kate, Lisa, Carol, Kim, Ann). On each of the 100 training trials, a ready signal appeared on the center of the CRT screen, and a word was subsequently played over the headphones. Subjects used the labelled keys on the monitor to enter the name of the correct talker. After all the subjects had entered their responses, or after three seconds, the correct name appeared on the screen. Responses entered after three seconds were recorded as failures to respond, and counted as incorrect.

After subjects completed two cycles of familiarization and training, they participated in a test phase. The test phase was identical to the training phase, except that subjects received no feedback indicating the correct talker for each token. After each day of training, subjects were given score cards indicating their percent correct for the two training phases and the test phase.

Test words were drawn from the same hundred word subset as those used in the familiarization and training phases, but there was no overlap of individual tokens between the test phase and the familiarization and training phases. The training stimuli were re-selected for each training day, so as to maximize variability in learning stimuli. On the tenth day of the experiment, trained subjects completed a final generalization test before they began the memory task. The generalization words were 50 new words, which the subjects had not heard previously, produced by the same ten talkers. That is, the words themselves, as well as the tokens, were novel. The generalization test was similar to the test phase on training days, but consisted of only 50 trials instead of a hundred.

Memory Test. Both trained and untrained subjects participated in the memory test. Subjects recalled a total of twelve ten-word lists: the first two were considered practice lists, and the remaining ten were scored for accuracy of serial recall. Words were presented at a rate of one every 1.5 seconds. A 500 ms 1000 Hz tone cued subjects when the word lists were about to begin, and also signalled the beginning and end of the recall period. Subjects were instructed to recall word lists in exact order of presentation, and to record their answers on separate response sheets. Subjects were given 60 seconds to recall the items between each list presentation. Half of the subjects heard lists produced by a single male talker, and half heard lists composed of tokens from each of the ten talkers.

Results and Discussion

Results of Training. Results of the training procedure are shown in Figure 1.

Insert Figure 1 about here

Average percent correct identification of talkers in the test phase of training is plotted across training days. Chance performance is set a twenty percent correct, assuming that subjects could minimally discriminate between male and female talkers, and limit the possible choices to gender appropriate talkers only. The results displayed in Figure 1 show that subjects consistently identified talkers above chance, and improved steadily across training days. An independent t-test confirmed that subjects were consistently operating at above chance, even after only 1 day of training [$t(16) = 12.08, p < .001$]. This pattern of results supports the hypothesis that training with voices helped subjects develop some stable, long-term memory representation for talker information.

Memory Results. Trained subjects were assigned randomly to groups, with the stipulation that identification performance for trained subjects should not vary significantly between conditions. An independent t-test for the means of training performance on the last day of training showed no differences between talker identification performance for trained subjects in the two talker conditions [$t(14) = -.27, p < .79$]. Subjects' responses on the memory test were scored for accuracy of serial recall. Responses were only scored as correct if subjects provided the correct word, or phonetic equivalent, in the exact serial position at which it occurred on the list. A four way analysis of variance (Training x Talker x Confusibility x Serial Position) revealed a significant effect for serial position [$F(9, 279) = 82.89, p < .05$], reflecting the characteristic shape of the serial recall curve. A significant main effect for talker was also found [$F(1, 31) = 4.34, p < .05$], replicating the results of Martin et al. (1989), Logan & Pisoni (1987), and Goldinger et al. (under review).

Figure 2 shows the pattern of recall across serial positions for multiple- and single-talker lists for the two groups of subjects.

Insert Figure 2 about here

The first panel shows the results for untrained subjects. At this relatively fast presentation rate, untrained subjects show the typical advantage for single-talker lists over

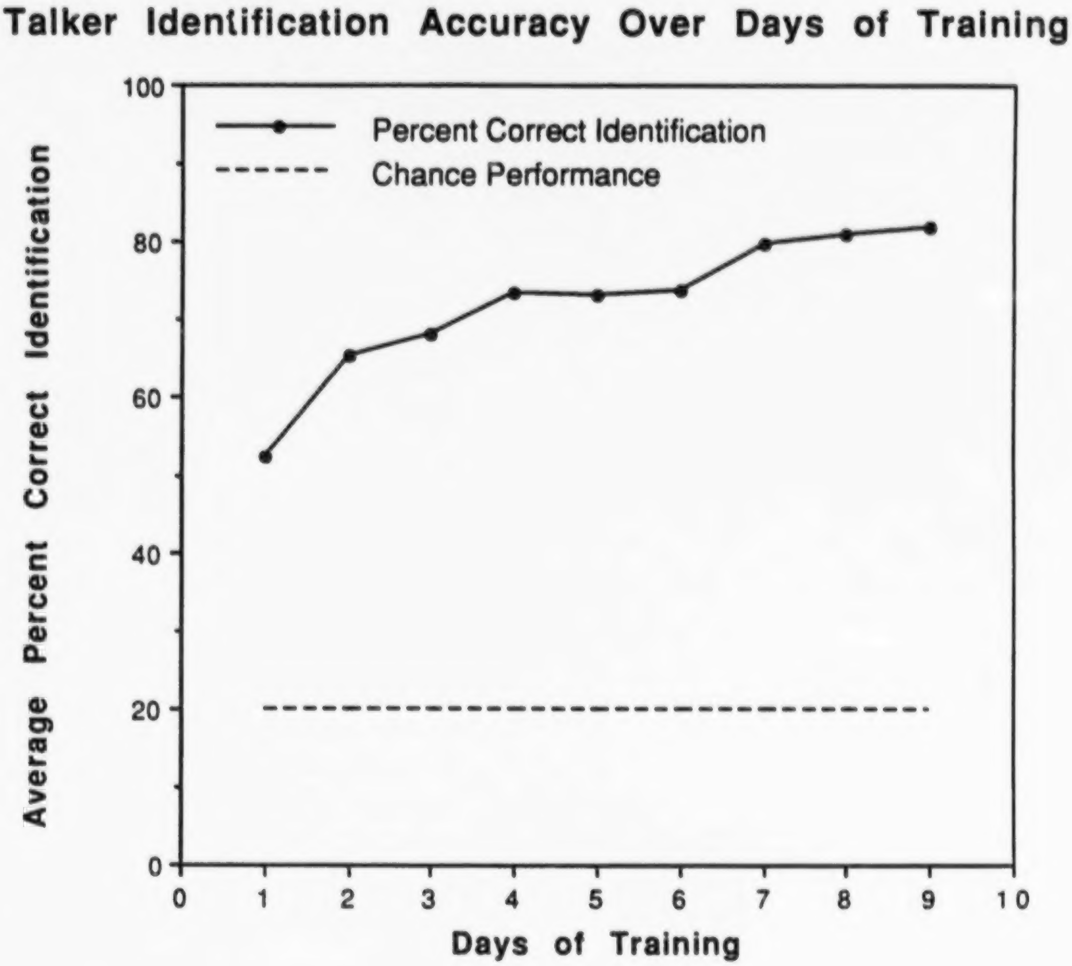
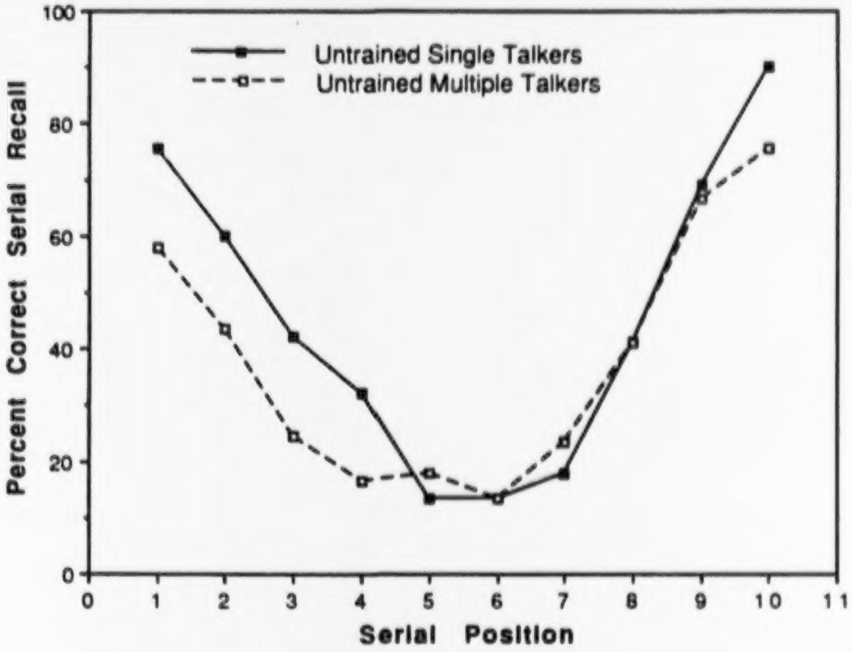


Figure 1. Average percent correct talker identification across nine days of training.

a) Multiple vs. Single Talkers -- Untrained Subjects, 1.5 Second ISI



b) Multiple vs. Single Talkers -- Trained Subjects, 1.5 Second ISI

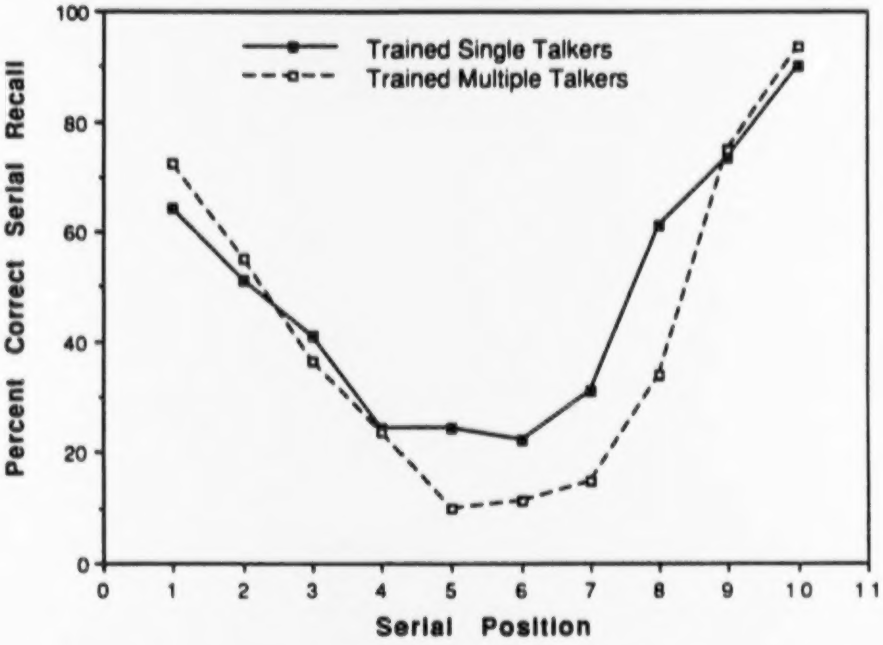


Figure 2. Percent correct serial recall for multiple and single talkers at two levels of training.

multiple-talker lists for serial recall in the primacy portion of the curve. The second panel of Figure 2 shows the recall results for trained subjects. For trained subjects, the recall advantage for single-talkers lists has been eliminated. In fact, serial recall for trained subjects shows a slight trend toward a multiple-talker advantage for primacy recall. Training subjects to encode and recall talker specific information seems mainly to have benefited recall for the multiple-talker lists: recall for single-talker lists is relatively unaffected by training. A siginificant three-way interaction between training, talker, and serial position, $[F(9,279) = 3.45, p < .01]$ also reflects this pattern of results. This cross-over advantage for trained subjects on multiple-talker lists in primacy is similar to that found by Goldinger et al. (under review) with slowed presentation rates. Apparently, by increasing accessibility of talker information, subjects are able to use talker-specific cues in serial recall from long-term memory, even at faster presentation rates. The main effect for training, however, did not reach significance $[F(1,31) = 2.03, p = .16]$. This may be due in part to the cross-over effect for training, including a slight decrement in single-talker recall for trained subjects.

The word-list confusibility factor also produced a significant main effect $[F(1,31) = 90.37, p < .01]$, replicating the results of Logan & Pisoni (1987) and Goldinger et al. (under review), and confirming the effectiveness of this manipulation. A significant interaction was also found for confusibility by serial position $[F(9,279) = 4.71, p < .01]$. As in Goldinger et al.'s study, confusibility had a greater effect in the primacy portion of the recall curve. This pattern of results can be seen in Figure 3, which shows recall for easy and hard word lists collapsed across talker and training condition.

Insert Figure 3 about here

This primacy-specific effect of word confusibility is assumed to reflect either decreases in rehearsal effeciency due to encoding lags for hard-to-recognize or low frequency words, or difficulty in retrieving low frequency words from memory (Goldinger et al., under review).

Two final, somewhat puzzling results were a significant two-way interaction between confusibility and talker $[F(1,31) = 6.07, p < .05]$ and a marginal three-way interaction between confusibility, training, and serial position $[F(9,279) = 1.86, p = .058]$. Figure 4 shows the pattern of results for confusibility in the two talker conditions, collapsed across training.

Insert Figure 4 about here

For easy words, multiple- and single-talker performance was comparable, but for hard word lists, multiple talkers did much worse than single talkers in the primacy portion of the

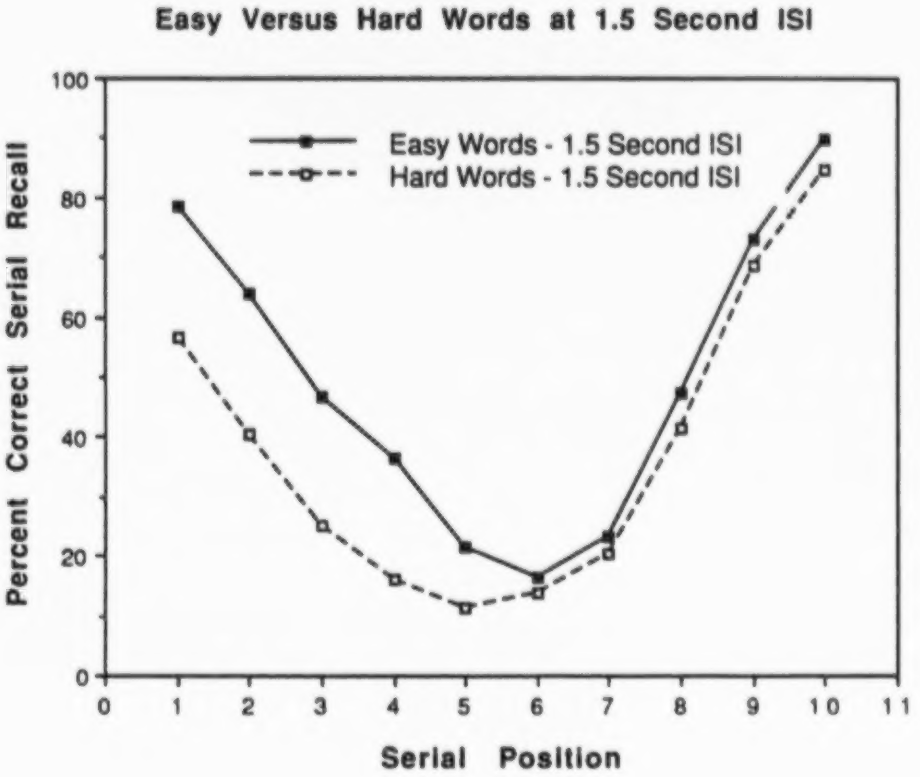
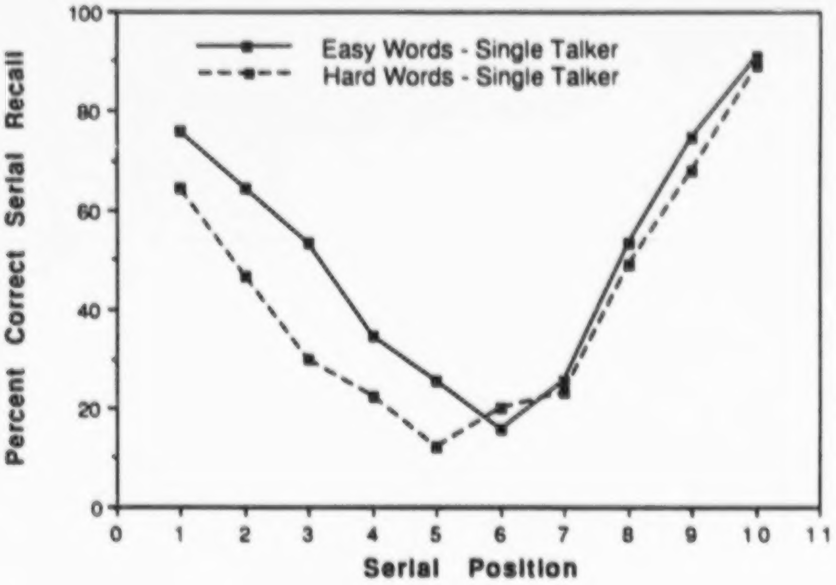


Figure 3. Percent correct serial recall for easy (non-confusible) versus hard (confusible) words.


a) Easy versus Hard Words -- Single Talker Lists

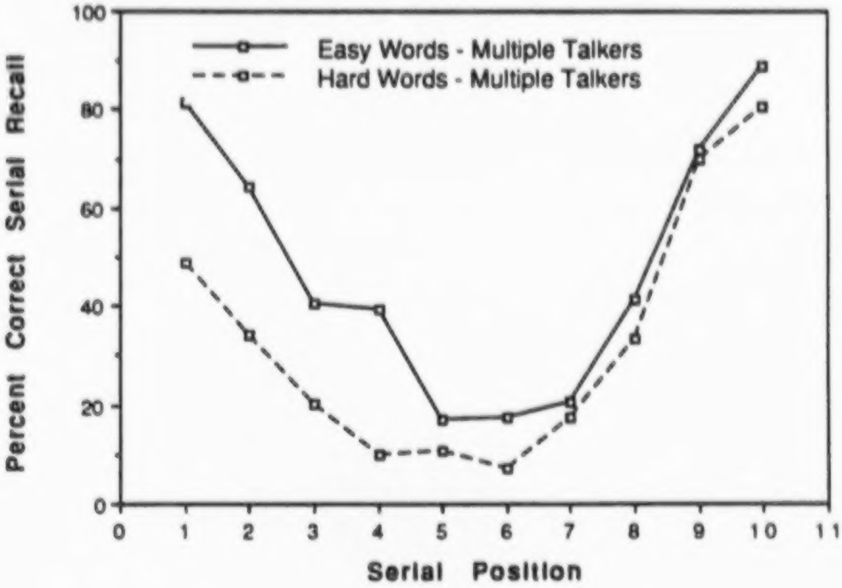

b) Easy versus Hard words -- Multiple Talker Lists

Figure 4. Percent correct serial recall for easy (non-confusible) versus hard (confusible) words for multiple- and single-talker word lists, collapsed across training.

recall curve. Because perceptual difficulty seems to have had a greater effect on multiple-talker lists than single-talker lists, this interaction suggests that encoding difficulties may be partially responsible for the deficits in recall for multiple-talker word lists found at very fast presentation rates. The interaction between talker and confusibility, however, has not been found previously in studies manipulating confusibility and multiple- versus single-talker recall (Martin et al., 1989; Logan & Pisoni, 1987; Goldinger et al., under review). Furthermore, if perceptual encoding deficits were responsible for primacy recall differences for multiple- and single-talker word lists, we would also expect to see a significant interaction between confusibility, talker, and serial position. The three way interaction between confusibility, talker, and serial position did not acheive significance $[F(9,279) = 1.42, p = .18]$.

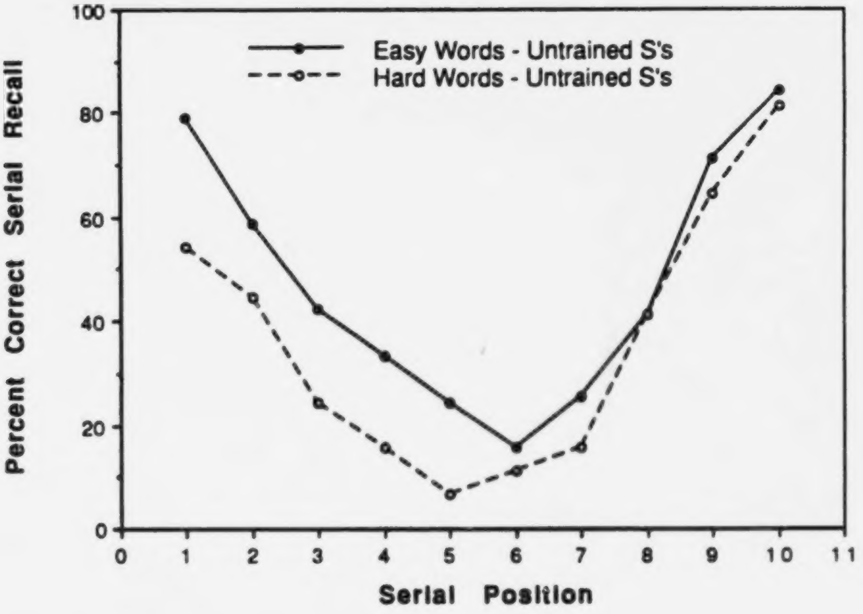
The interaction between confusibility, training, and serial position can be interpreted by examining Figure 5, which shows recall for easy and hard words at two levels of training.

Insert Figure 5 about here

The difference between easy and hard words is larger for trained subjects than for untrained subjects. Again, the differences are restricted to the primacy region of the curve. Trained subjects appear to do slightly better on easy word lists and slightly worse on hard word lists than untrained subjects. It is unclear how to explain these result, unless the salience of highly familiar talker information leads to greater perceptual deficits for familiar voices. This explanation is highly tentative and post hoc, and probably incorrect given the marginal nature of the result and the small number of subjects run. It is also important to note that training, if anything, makes encoding more difficult for multiple-talker stimuli. It is highly unlikely then, that training produces an increase in multiple-talker recall in primacy by easing encoding of words produced by multiple talkers.

The results of Experiment 1 show improvement in primacy recall for multiple-talker word lists for highly familiar voices, even at relatively fast presentation rates. This effect is unlikely to be due to automatization of perceptual encoding of multiple-talker stimuli: if anything, training seemed to increase the difficulty of encoding confusible words. When subjects are trained to identify voices, recall for multiple-talker word lists improves to the level of recall for single-talker word lists, even at a relatively fast presentation rate. These results suggest that further increasing accessibility of talker-specific cues by slowing presentation rate may lead to a multiple-talker advantage for recall of words produced by familiar talkers. In Experiment 2, this hypothesis was explored directly by slowing down the presentation rate for trained subects from a 1.5 ISI to a 4 second ISI. At this slower rate, highly familiar voice cues should lead to greatly improved recall for multiple-talker word lists in the primacy region of the serial recall curve.

a) Easy versus Hard words -- Untrained Subjects



b) Easy versus Hard words -- Trained Subjects

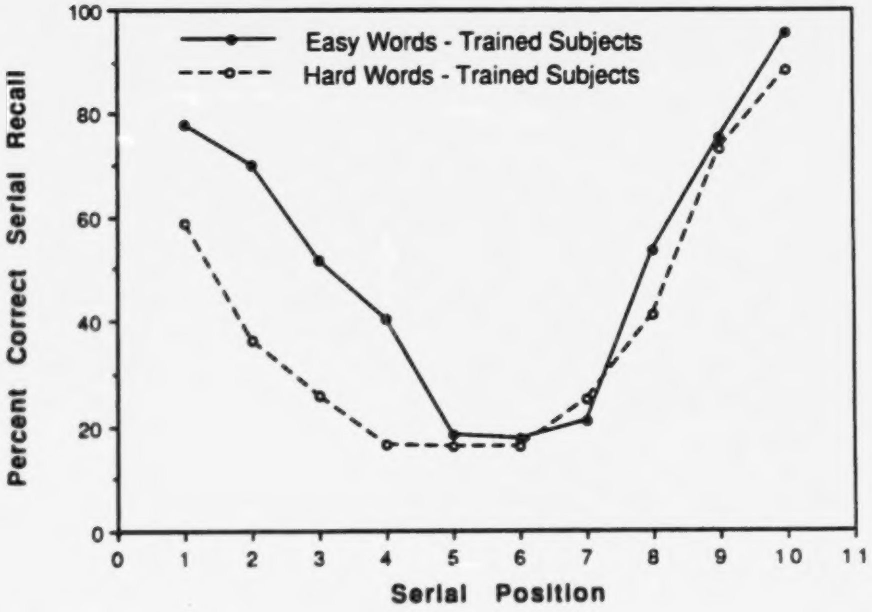


Figure 5. Percent correct serial recall for easy (non-confusible) versus hard (confusible) words at two levels of training.

Experiment 2

Method

Subjects. The same 17 subjects that participated in the training condition of Experiment 1 served as subjects in Experiment 2. One subject dropped out because of illness. Subjects were paid for their participation.

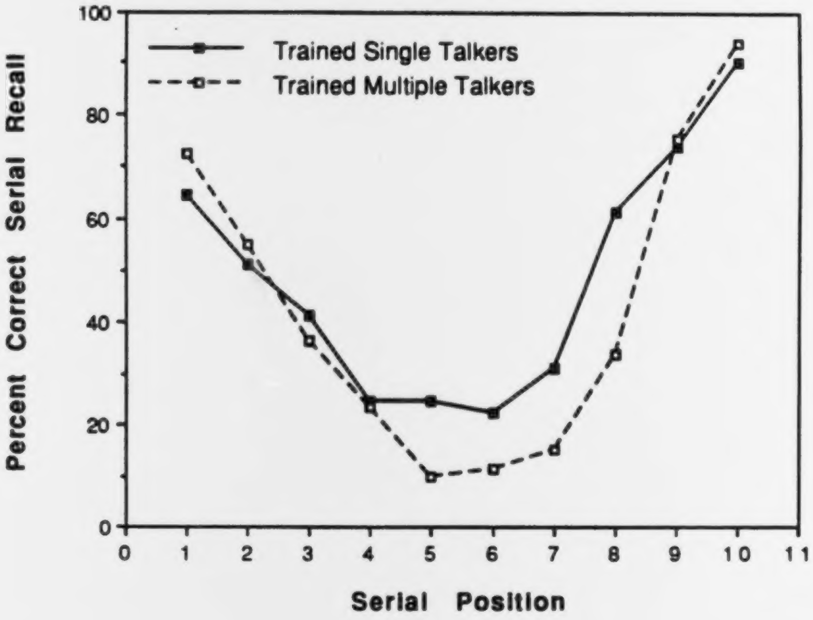
Materials and Procedure. The stimuli and procedures for the memory test were identical to those for Experiment 1, with one exception. The presentation rate in the memory experiment was slowed from a 1.5 second ISI to a 4 second ISI. Prior to the memory experiment, subjects participated in a review training session, which lasted approximately an hour. They repeated a full day of training, except that instead of the final 100-trial identification test, they received a 50-trial generalization test. After a five to ten minute break, subjects then completed the memory test.

Results and Discussion

Subjects showed evidence of retention for long-term memory representation of talker information. The mean percent correct identification in the generalization test was well above chance [t(12) = .55, p < .59]. Memory protocols were scored for serial recall as in Experiment 1. Figure 6 shows recall for trained subjects on single- and multiple-talker lists at two presentation rates: the trained-subject data from Experiment 1 versus the data from Experiment 2.

Insert Figure 6 about here

A three-way analysis of variance (Talker x Confusibility x Serial Position) on the data for Experiment 2 showed a significant main effect for serial position [$F(9,126) = 16.6, p < .01$], but no main effect for talker [$F(1,14) = .26, p < .62$]. This null result for the talker manipulation is surprising given the robustness of the differences between single and multiple talkers in serial recall found by Martin et al., (1989), Logan & Pisoni (1987), and Goldinger et al. (under review). This may be accounted for, however, by the cross-over effect for serial recall between the two talker conditions displayed in the lower panel of Figure 6. Multiple-talker lists do show the expected large advantage for serial recall in the first two list positions. In the remaining list positions, however, there is a recall advantage for single-talker lists. A marginal serial position by talker interaction supports this interpretation of the data [$F(9,126) = 1.74, p < .09$].


a) Trained Subjects at 1.5 Second ISI

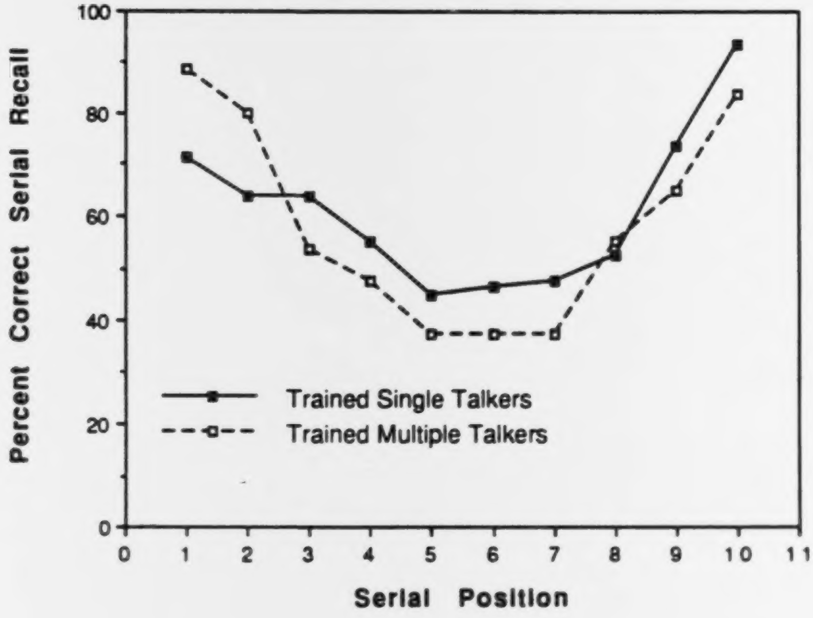

b) Trained Subjects at 4 Second ISI

Figure 6. Percent correct serial recall for trained subjects at two presentation rates.

As in Experiment 1, the analysis of variance revealed a significant main effect for confusibility [$F(1,14) = 14.5, p < .01$], and a significant two-way interaction between confusibility and serial position [$F(9,126) = 3.0, p < .01$]. Again, these results reflect a greater advantage for easy words over hard words in early list positions. The confusibility by serial position by talker interaction was not significant [$F(9,126) = .61, p < .78$]. This replicates earlier findings (Martin et al., 1989, Logan & Pisoni, 1987; Goldinger et al., under review), and again fails to support the hypothesis that encoding time lags are responsible for talker differences in primacy.

General Discussion

The present experiments were designed to further examine the role of encoding differences for single- and multiple-talker word lists in serial recall. Subjects were trained to become familiar with talker-specific cues in order to test two hypotheses. One account of the multiple and single talker differences in recall states that subjects mandatorily rehearse talker-specific cues, leading to worse recall for multiple talkers at fast presentation rates, but a multiple-talker advantage at long presentation rates, where talker specific cues can aid in retrieval from long-term memory. Under this hypothesis, training subjects to recognize and use talker-specific information should lead to enhanced serial recall in the primacy portion of the recall curve, even at faster presentation rates. Furthermore, training should interact with both talker condition and serial position, reflecting differences in rehearsal efficiency and ease of retrieval from long-term memory.

An alternative explanation for differences in recall between multiple- and single-talker word lists is that normalizing for talker differences in multiple-talker word lists increases the difficulty of encoding, leading to a time lag in processing, which decreases efficiency of rehearsal and transfer to long-term memory. Under this hypothesis, training subjects to recognize voices was predicted to ease the process of talker normalization, again leading to increased efficiency of processing, and increased recall for multiple-talker word lists in the primacy portion of the serial recall curve. Training would also be expected to interact with both talker condition and confusibility under this hypothesis, reflecting the role of encoding difficulty in serial recall for single and multiple talkers.

The results support an explanation based on rehearsal and elaboration of enriched stimulus information for multiple-talker stimuli. Multiple-talker lists showed improvement in primacy recall with training, and began to surpass single-talker lists at a relatively fast presentation rate. When presentation rate was slowed, this multiple-talker primacy advantage for trained subjects appeared to increase even further, although direct statistical comparisons could not be made between the first and second experiments. Previous research indicates that subjects do develop incidental memory representations of talker-specific information (Geiselman & Belezza, 1977), and that the processing of this voice information may be responsible for differences in recall for multiple- and single-talker word lists (Goldinger et al., under review). Apparently, training subjects to develop stable and accessible memory representations of voice information eases either rehearsal or retrieval of these talker cues in a serial recall task.

The results are more ambiguous with regard to the second hypothesis. If training acted to decrease encoding time for multiple-talker lists, we would expect to see an interaction between training condition and both talker condition and list confusibility. Neither of these results was obtained. A significant interaction was found, however, between the confusibility factor and talker condition, indicating a possible role for encoding factors in recall differences for multiple and single talkers. This could be an indication that encoding efficiency is also in part responsible for recall differences for multiple and single talkers.

In an earlier study from our laboratory, Luce, Feustel, and Pisoni (1983), for example, found depressed recall for synthetic speech word lists when compared with natural speech word lists. Intuitively, it seems unlikely that subjects would be more likely to rehearse talker information for synthetic speech as opposed to natural speech, unless attention is directed to rehearsing voice information only when it deviates from a norm or expectation. More importantly, recall for synthetic word and natural word lists failed to interact with presentation rate in this study. Unfortunately, Luce et al. manipulated rate in a free recall task, as opposed to a serial recall task, making direct comparisons with Goldinger et al.'s results difficult. Changes in presentation rate have been shown to affect rehearsal processes in both serial and free recall, however, making a rehearsal-based explanation of Luce et al.'s findings somewhat implausible. Deficits in serial recall in the primacy portion of the serial position curve may be due in part to perceptual deficits as well.

There is the additional possibility, however, that the interaction between confusibility and talker condition simply represent a sampling error. A number of previous studies have failed to find an interaction between confusibility and talker condition (Martin et al., 1989; Logan & Pisoni; 1987; Goldinger et al., under review). The pattern of recall for easy and hard word lists at two levels of training indicates that, if anything, training enhanced the difficulty of encoding relative to untrained talkers. Furthermore, there was no interaction between confusibility, talker condition, and training, yet training did improve multiple-talker recall in the primacy region of the curve.

Given that deficits associated with multiple-talker stimuli have been demonstrated to interact with low level perceptual processes, it is reasonable to assume that these perceptual deficits may lead to differences in rehearsal efficiency and recall as well. This hypothesis is not, however, incompatible with the finding that talker cues are processed integrally with phonetic and semantic information (Mullennix & Pisoni, 1988; Goldinger et al., under review). Taken together, the results of Experiments 1 and 2 provide support for Goldinger et al.'s hypothesis that mandatory rehearsal of multiple-talker cues accounts at least in part, for differences in multiple- and single-talker recall. The role that perceptual deficits play in these recall differences, however, remains less clear.

In summary, the present investigation has shown that training subjects to use and remember voice-specific information improves multiple-talker recall in the primacy portion of the serial recall curve even at a relatively fast presentation rate. When the presentation rate is slowed, trained subjects show an advantage for recall of multiple-talker lists, at least in early list positions. These results support the conclusion that talker-specific cues are processed mandatorily in speech perception. In the case where the subject has no long term representation of voice information, or where processing demands are too high to allow for complete processing and transfer to long term memory, it is proposed that processing of complex talker information leads to recall deficits in the primacy region of the serial recall curve. If talker specific cues are more accessible, however, or if subjects are allowed sufficient time for processing and transfer of talker-specific information to long term memory, unique talker-specific cues may lead to greater elaboration of the item during rehearsal or to the formation of more effective retrieval cues for serial recall (Martin et al. 1989; Logan & Pisoni, 1987; Goldinger et al., under review).

References

Atkinson, R.C., & Shiffrin, R.M., (1968). Human memory: A proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.) *The psychology of learning and motivation*, **2**, 89-105. New York: Academic Press.

Bjork, R.A., & Whitten, W.B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, **6**, 173-189.

Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, **26**, 1-7.

Cowan, N., (1988). Evolving concepts of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, **104**, No. 2, 163-191.

Cowan, N., (1984). On short and long auditory stores. *Psychological Bulletin*, **96** (2), 341-370.

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.

Craik, F.I.M, & Lockhart, R.S. (1972). Levels of precessing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, **11**, 671-684.

Crowder, R.G., (1976). *Principles of learning and memory*. Erlbaum: Hillsdale, NJ.

Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, **29**, 655.

Geiselman, R.E., & Bellezza, F.S. (1976). Long-term memory for speaker's voice and source location. *Memory and Cognition*, **4**, 483-489.

Glanzer, M., & Cunitz, A.R (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, **5**, 351-360.

Glenberg, A.M., Bradley, M.M., Kraus, T.A., & Renzaglia, G.J. (1983). Studies of the long-term recency effect: Support for a contextually guided retrieval hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9**, 231-255.

Glenberg, A.M., Bradley, M.M., Stevenson, J.A., Kraus, T.A., Tkachuk, M.J., Gretz, A.L., Fish, J.H., & Turpin, B.M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 475-479.

Glenberg, A.M., & Swanson, N.G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**, 3-15.

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (under review). On the locus of talker variability effects in recall of spoken word lists.

Greene, R.L. (1986). Sources of recency effects in free recall. *Psychological Bulletin*, **99** (2), 221-228.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **29**, 98-104.

Logan, J.S., & Pisoni, D.B. (1987). Talker variability and the recall of spoken word lists: A replication and extension. *Research on speech perception progress report no. 13*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25** (1), 17-32.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 676–684.

Mattingly, I.G., Studdert-Kennedy, M., & Magen, H. (1983). Phonological short-term memory preserves phonetic detail. *Journal of the Acoustical Society of America*, **73**, 56.

Modigliani, V., & Hedges, G.H. (1987). Distributed rehearsals and the primacy effect in single-trial free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 426-436.

Mullennix, J.W., & Pisoni, D.B. (1988). Detailing the nature of talker normalization in speech perception. *Research on speech perception progress report no. 14*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Mullennix, J.W., & Pisoni, D.B. (in press). Talker variability effects and processing dependencies between word and voice. *Perception & Psychophysics*.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability of spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365–378.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception, Progress Report No. 10*, Speech Research Laboratory, Psychology Department, Indiana University, Bloomington, Indiana.

Pisoni, D.B. and Luce, P. (1987). Acoustic-phonetic representations in spoken word recognition. *Cognition*, **25**, 21–52.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, **89**, 63-77.

Shimizu, H. (1987). The relationship between memory performance and number of rehearsals in free recall. *Memory & Cognition*, **15**, 141-147.

Sumby, W.H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, **1**, 443-450.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, **60**, 198-212.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Inhibition or Facilitation? An Investigation of Form-based Priming and Response Bias in Spoken Word Recognition[1]

Joanne K. Marcario and Stephen D. Goldinger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This research was supported by NIH Research Grant NS-12179-12 to Indiana University, Bloomington, IN. We thank Paul Luce and David Pisoni for helpful suggestions and criticisms.

Abstract

Two recently reported experiments employing similar auditory priming paradigms to study spoken word recognition have yielded opposite results. In one study, Slowiaczek, Nusbaum and Pisoni (1987) observed facilitation of target word identification when related prime-target pairs shared common phonemes. In a second study, however, Goldinger, Luce and Pisoni (1989) observed inhibition of target word identification when related prime-target pairs were phonetically confusable, but shared no common phonemes. The present investigation replicated the major results of these two experiments using the same priming task, the same target words, and the same unrelated primes. The only factor that varied was the degree of phonetic overlap between the related prime and its target. The results supported the previous findings: the replication of the Slowiaczek et al. conditions yielded facilitation of target identification, and the replication of the Goldinger et al. conditions yielded inhibition of target identificaion. However, by utilizing a common pool of *unrelated* prime-target pairs across both experimental contexts, we could compare the experiments more directly. The present findings suggest that the facilitation reported in the Slowiaczek et al. (1987) study may be due to expectancies or biases generated by the subjects after continued presentation of prime-target pairs that contain identical word-initial phonemes. The inhibition obtained by Goldinger, et al. appears to be due to the phonetic confusability between the related primes and their targets, not expectancies generated by the subject. The results are discussed in terms of the role of expectancies and selective attention in the priming paradigm.

Inhibition or Facilitation? An Investigation of Form-based Priming and Response Bias in Spoken Word Recognition

A common methodology employed in the investigation of spoken word recognition is the auditory priming paradigm. A priming task typically consists of the presentation of a test word, called a *target* that is preceded by another word, called a *prime*. Typically, following target presentation, subjects are required to either identify or execute some speeded response to the target. The independent manipulation of interest in priming experiments is the particular relations shared between primes and targets. For example, primes may share acoustic-phonetic information (e.g., *bat - bill*) or knowledge-based information (e.g., *cat - mouse*) with the targets. Primes may also be totally unrelated to the targets, and thereby serve as a control against which subjects' performance on related prime-target pairs may be evaluated. The dependent variable of interest in priming tasks is usually either the subjects' reaction times or percentages of correct identification of the targets. Recognition of primed targets can be compared to unprimed targets, or to targets paired with unrelated primes, to assess the magnitude of any effects the primes exert on the identification of the targets.

Typically, priming manipulations yield facilitation of responses to primed targets, relative to unprimed targets. Indeed, Ratcliff and McKoon (1978) *define* a semantic priming task as "... the facilitation of response to one test item as preceded by another." There are numerous examples in the literature of experiments demonstrating the facilitatory effects of priming targets by visually presenting semantically-related words (Martin & Jansen, 1988; Collins & Loftus, 1975;). Similarly, facilitatory priming effects have been observed in experiments using visually-presented phonologically-related words as primes and targets (Hillinger, 1980), and in priming of visually-presented targets by spoken, phonologically-related primes (Jakimik, Cole & Rudnicky, 1985). Although less research has been conducted to investigate the effects of priming on recognition of spoken words than of printed words, facilitatory priming for spoken primes and targets has been reported in the literature by Slowiaczek, Nusbaum, and Pisoni (1987).

Although the facilitation of target recognition in both semantic and form-based priming experiments is well-established (e.g. Collins & Loftus, 1975; Neely, 1977), some recent studies have demonstrated *inhibition* of target recognition, using both visual (Meyer, Schvaneveldt & Ruddy, 1974; Neely, Schmidt & Roediger, 1983; Taraban & McClelland, 1987), and auditory priming tasks (Tanenhaus, Flanigan & Seidenberg, 1980; Slowiaczek & Pisoni, 1986). To date, few cohesive accounts have been offered for both the inhibitory and facilitatory effects of priming. A single explanation that encompasses both priming effects may not be possible; it may be the case that qualitatively different processes underlie the inhibition and facilitation of target identification produced by priming, despite the apparent methodological similarities across tasks. Our understanding of the processes involved in priming may be improved if we consider the reported findings in light of theories of spoken word recognition and lexical access. Two contemporary theories will be discussed, in order to examine their postulated processes and consider how they may account for both inhibitory and facilitatory priming effects.

Cohort theory (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987) is perhaps the most prominent theory in the spoken word recognition literature. Cohort theory assumes a bottom-up basis of lexical activation, and subsequent interactions with top-down syntactic and semantic information that allow listeners to identify words quickly and accurately. In cohort theory, word recognition proceeds in the following way: The initial acoustic-phonetic information provided by the speech waveform activates a "cohort" of all the words in the lexicon with the same initial sound sequence. As the speech input proceeds, word candidates are eliminated from the original cohort (e.g., deactivated) by an interaction of top-down contextual information and mismatched acoustic-phonetic information until all but one word, the "recognized" word, remains. Cohort theory assumes that as words are processed in real time, more weight is given to the beginnings of words than to the endings of words. In a phonological priming task, therefore, the residual activation from the recognition of the prime word should pre-activate the initial cohort of the target word. Given this assumption, cohort theory predicts facilitation in a task in which the prime shares word-initial phonological information with the target word.

Slowiaczek, Nusbaum, and Pisoni (1987) conducted a series of three experiments using an auditory priming paradigm to test this prediction of cohort theory. The first two experiments used related prime-target pairs, in which either real-word primes or pseudoword primes shared zero, one, two, three, or all phonemes in common with the target words. According to cohort theory, facilitation of target identification should increase as the number of word-initial phonemes shared by primes and targets increases. This is precisely the result Slowiaczek et al. obtained in both experiments, although the effects were attenuated in the experiment with the pseudoword primes. In addition, facilitation of target identification increased as the target stimuli were increasingly degraded by noise. These results support cohort theory's assumptions that words are recognized sequentially, from "left to right", in accordance with their temporal presentation, and that activation of the components of the word-initial cohort may occur via bottom-up, acoustic-phonetic input.

Despite the support for cohort theory implicit in the results of their first two experiments, the results of the third experiment conducted by Slowiaczek et al. *contradict* cohort theory's prediction that facilitation of target identification should occur only if the primes and targets share word-initial phonemes. In the third experiment, primes were related to the target words by the number of phonemes shared from the ends of the words. Contrary to cohort theory's prediction, Slowiaczek et al. found that facilitation of target identification also increased as the number of word-final phonemes shared by primes and targets increased. Since an initial cohort is activated by the first phoneme of a word, cohort theory predicts that primes and targets that do not begin with the same sounds should have no residual effects on each other. Accordingly, the results of Slowiaczek et al.'s third experiment calls the directionality of Marslen-Wilson's proposed cohorts into question.

Another theory of word recognition, proposed recently by Luce and Pisoni (1987), is the ***neighborhood activation model*** (NAM) of spoken word recognition. NAM describes the recognition of words as a function of ***similarity neighborhoods***, which are comparable to the cohorts of word candidates described by Marslen-Wilson. Similarity neighborhoods are defined as collections of words in memory which are phonetically similar to any given referent words. In NAM, however, the importance of phonetic similarity in lexical processing is not limited to the initial segments of words. Instead, in NAM, the similarity neighborhood for any given word is determined by global phonetic similarity between the word and its neighbors, not exclusively by word-initial similarity.

Two structural characteristics of similarity neighborhoods have been shown to affect word recognition speed and accuracy: ***neighborhood density*** and ***neighborhood frequency***. Neighborhood density refers to the absolute number of words in a similarity neighborhood; sparse neighborhoods contain few words, and dense neighborhoods contain many words. Neighborhood frequency refers to the average frequency of all words in a similarity neighborhood; high frequency neighborhoods contain mostly high frequency words, and low frequency neighborhoods contain mostly low frequency words.

In NAM, word recognition is hypothesized to be a two-stage process. Upon the presentation of a stimulus word, a similarity neighborhood of all acoustically similar words is activated in long-term memory. Once the neighborhood has been activated, ***word decision units*** function to discriminate the target word from its activated neighbors. According to NAM, increased activation of a target word's neighbors should decrease the probability of identifying the target word itself. One way of manipulating the level of activation of a neighborhood is by means of a priming task. If the stimulus word is primed with a phonetically related word, the residual activation created by the related prime should produce neighborhood competition for recognition (Goldinger, Luce, & Pisoni, 1989). Thus, NAM predicts ***inhibition*** of target word identification in form-based priming tasks.

NAM is formally represented by the *neighborhood probability rule*, which is based on R.D. Luce's (1959) choice rule, and has the form:

$$p(ID) = \frac{p(StimulusWord) * Freq_s}{p(StimulusWord) * Freq_s + \sum_{j=1}^{n} [p(Neighbor_j) * Freq_j]}$$

in which ***p(stimulus word)*** is the probability of the stimulus word, $\boldsymbol{freq_s}$ is the frequency of the stimulus word, $p(neighbor_j)$ is the probability of neighbor j, and $freq_j$ is the frequency of neighbor j.

The rule states that the probability of correctly identifying a target word is equal to the probability of the target word divided by the probability of the target plus the combined probabilities of the target word's neighbors[1], given neighborhood density, neighborhood frequency and stimulus word frequency. Neighborhood density is represented in the denominator of the target as the summed probabilities of all the target word's neighbors. Frequency is used as a weighting function on the probabilities of the stimulus word and its neighbors, biasing an identification decision in favor of higher frequency words. The neighborhood probability rule predicts that priming with a phonetically related neighbor should increase the summed neighbor probability term in the denominator of the rule, thereby decreasing the probability of target identification.[2]

Goldinger, Luce, and Pisoni (1989) conducted two experiments to test two of the predictions that NAM makes regarding form-based priming effects in spoken word recognition. The first prediction was that phonetically related primes would inhibit target identification because of increased neighborhood competition. The second prediction was that low frequency primes would produce relatively *more* inhibition than high frequency primes. The results of both experiments supported the predictions of NAM. Goldinger et al. found that priming with phonetically related words significantly inhibited target recognition, that words from sparse neighborhoods were identified more accurately than words from dense neighborhoods, and that high frequency words were identified more accurately than low frequency words. Furthermore, the three conditions in which the priming effect reached statistical significance all contained low frequency primes. This finding demonstrated that low frequency primes produce stronger inhibition of target word identification than high frequency primes.

The findings reported by Slowiaczek et al. (1987) and by Goldinger et al. (1989) show opposite effects of priming in apparently very similar experimental situations. The tasks used in each set of experiments were very similar, differing primarily in the degree and type of phonological overlap between the related primes and their targets. Whereas Slowiaczek et al. selected prime-target pairs that shared common phonemes, Goldinger et al. selected prime-target pairs that were phonetically *similar* but shared no common phonemes. Conflicting findings such as these warrant further investigation.

Goldinger et. al used prime-target pairs that shared no common phonemes in an attempt to dissuade subjects from developing response strategies. A number of researchers have suggested that facilitation in priming tasks may be due to expectancies or biases generated by subjects, because of the high degree of salient similarity or association between the primes and their respective targets. Becker and Killion (1977), for example, proposed that, via priming, an experimenter can manipulate the expectancies of the subject such that "If subjects can be induced to expect one of a small set of stimuli, to the exclusion of others, then the expected stimulus may benefit from the bypassing of feature extraction, whereas the unexpected stimuli may not" (p. 400). In a similar vein, Posner and Snyder (1975) state that in their matching and classification experiments, "...attention to the prime often appears to be used to match the prime item against the array [target]. This serves to facilitate the 'yes' responses to matching pairs when the prime matches the array and 'no' responses to mismatching prime-array pairs. This strategy alone can account for many of our results..." (p.680).

---

[1] The expressions "probability of the target" and "probabilities of neighbors" refer to confusion matrix estimates of the intelligibility of the separate segments of the words in question for our particular speaker's voice, at several signal-to-noise ratios. See Luce (1986) for details.

[2] It should be stressed that this prediction of the neighborhood probability rule is assumed only for cases of *form-based* priming, and is not assumed without modification for cases of semantic, or knowledge-based, priming.

The present experiments were conducted to determine the reason or reasons for the conflicting results of the Goldinger, et al. and Slowiaczek et al. studies. Specifically, the study was designed to allow for direct comparison of the experiments, in order to determine whether both experiments demonstrated "true" activation-based priming effects, as opposed to bias effects. Modified replications of both of the original studies were conducted. First, the studies were rendered comparable by using identical experimental procedures, identical unrelated primes and identical targets. New stimuli were generated for each experiment; the replication of the Slowiaczek et al. (1987) experiment utilized related primes which were chosen so that they were neighbors to the targets, and contained one word-initial overlapping phoneme (e.g. *bull* and *beer*). In the replication of the Goldinger et al. (1989) study, prime-target pairs related only by phonetic similarity were generated. Primes were chosen to be the nearest neighbors of the target words that shared no common phonemes. Thus, for example, *bull* and *veer* are considered related because they have a high probability of confusion of individual phonemes within the pair, yet they do not share any identical phonemes. In addition to the related primes-target pairs, unrelated prime-target pairs were also generated for both experiments so that a baseline could be obtained, against which the priming effects could be evaluated. Neutral prime-target pairs had a confusability rating of approximately zero and did not share any identical phonemes.

Across both replications, then, the prime-target pairs were either phonetically unrelated, only phonetically related, or phonemically related. As both experiments are essentially direct replications of previously published work, we predicted that in the replication of the Slowiaczek et al. (1987) study we would observe facilitation of the recognition of primed targets, and that in the replication of the Goldinger et al. (1989) study we would observe inhibition of the recognition of primed targets. However, by replicating the two experiments with a common set of stimulus materials, a critical comparison may be performed. The common element across both replications performed here is the selection of *unrelated* prime-target pairs. Although the unrelated pairs are typically selected to serve only as a control condition that may be used to determine the magnitude and direction of a priming effect, in the present study the unrelated pairs are actually of primary interest. Specifically, we may examine the *nature* of subjects' incorrect guesses on the unrelated prime-target trials across different experimental contexts. If subjects in one context or the other are employing a bias in selecting their responses, noticeable differences should be apparent in either the number of incorrect responses to the unrelated pairs, or in the nature of the specific errors committed.

Experiment 1

Method

Subjects. Forty-four Indiana University undergraduates participated as partial fulfillment of requirements of an introductory psychology course. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimuli. One-hundred and sixty-eight phonetically related prime-target pairs were selected from a computerized database based on Webster's Pocket Dictionary (1967). In addition, neutral primes were selected for each of the 168 targets, for a total of 504 words. The related prime-target pairings were created by searching the database for each target's nearest neighbor with no common phonemes. Degree of similarity of a given prime to its target word was computed using confusion matrices for individual consonants and vowels (see Luce, 1986, for a complete description). The neutral primes were selected by searching for words from neighborhoods that had approximately the same density as their prospective targets, but were not phonetically confusable with the targets.

From the original lists of words generated by these searches, the final 504 words selected were those which met the following constraints: (1) All targets and neutral primes were three phonemes in length; related primes were either two or three phonemes in length; (2) all words were monosyllabic; (3) all words were listed in the Kučera and Francis (1967) corpus; and (4) all words had a rated familiarity of 6.0 or above on a seven-point scale. These familiarity ratings were obtained from a previous study by Nusbaum, Pisoni, and Davis (1984). In that study, all the words from Webster's Pocket Dictionary were presented visually to subjects for familiarity ratings. The rating scale ranged from (1) "don't know the word" to (4) "recognize the word, but don't know its meaning" to (7) "know the word and its meaning." The rating criterion of 6.0 and above was used to ensure that all prime and target words would be known by the subjects.

All stimuli were recorded in a sound-attenuated booth by a male talker with a midwestern dialect. All words were spoken in isolation. The stimuli were then low-pass filtered at 4.8 kHz and digitized at a sampling rate of 10 kHz using a 12-bit analog-to-digital converter. All words were excised from the list using a digitally controlled speech waveform editor (WAVES) on a PDP 11/34 computer (Luce and Carrell, 1981). Finally, all words were paired with their appropriate counterparts and stored digitally as stimulus files on computer disk for later presentation to subjects during the experiment.

To ensure that all stimuli could be identified accurately, an additional group of subjects was asked to identify all the words in the clear. Words which were not correctly identified by at least 8 of 10 subjects were re-recorded and replaced.

Design. After selection constraints were satisfied, the sets of primes and targets were divided into four cells constructed by combining two levels of each of two variables: (1) target frequency and (2) neighboorhood density. In each of the four conditions, low frequency related and unrelated primes were used. Only low frequency primes were selected for the present experiments because Goldinger et al. (1989) found that only low frequency primes produced significant inhibitory priming effects. Once the prime-target pairs were assigned to their proper conditions, all conditions contained 42 pairs. Because every target item had two corresponding primes and no subject was to be presented the same target item twice, the stimuli were divided into two lists. Every subject responded to all 168 targets, but the related and unrelated primes varied across groups of subjects. For a given group, 84 targets were primed by related primes, 84 targets by control primes. An equal number of subjects were presented with each list. In this manner, all subjects were presented all targets, but the primes associated with those targets varied across groups. The dependent measure in each condition was the mean percentage of correctly identified targets.

Procedure. Subjects were tested in groups of six or fewer. Each subject was seated in a testing booth equipped with an ADM computer terminal and a pair of TDH-39 headphones. The presentation of stimuli was controlled by a PDP 11/34 computer. All stimuli were presented in random order.

A prompt appeared on the CRT screen saying, "GET READY FOR NEXT TRIAL." Five hundred milliseconds after the prompt appeared, a prime was presented over headphones at 75 dB (SPL) in the clear. Immediately upon the offset of the prime, 70 dB of white noise was presented. Fifty milliseconds after the presentation of the noise, the target item was presented at 75 dB (SPL), yielding a +5 dB signal-to-noise (S/N) ratio. The subject's task was to identify each target word and type the response on the ADM keyboard as accurately as possible following each trial. Subjects were not under a time constraint for responding.

After completion of the task, a questionnaire was administered to subjects. This questionnaire was designed to find out exactly what types of information the subjects were using to identify the target words in noise and to investigate the types of strategies subjects may have employed in response selection. There were three main questions included: (1) Were there any characteristics of the primes that helped the subject identify the targets? (2) In any of the word pairs, did the subjects notice identical speech sounds produced in both the prime and the target? (3) Did the subjects consciously make the task easier for themselves by using any type of strategy? Also, four questions about the experimental procedure but not about any strategy were included to allow the experimenters to assess the reliability of the data in each subject's questionnaire and to mask the motivation of administering the questionnaire as well. (See Appendix A for complete questionnaire.)

Results and Discussion

The percentage of words correctly identified was determined for each subject. For a response to be considered correct, the entire response either had to match the target item exactly, or had to be a homophone of the target word (e.g. *ate*, *eight*). All simple spelling or typing errors, such as letter transpositions, were corrected prior to data analysis.

Figure 1 displays the results of the priming manipulation in Experiment 1. Light bars show performance for targets preceded by related primes, dark bars show performance for targets preceded by unrelated primes. Three main effects and one significant interaction are displayed. Main effects were observed for prime type, target frequency, and neighborhood density, and a significant interaction was observed between target frequency and neighborhood density.

Insert Figure 1 about here

A three-way analysis of variance (prime type X neighborhood density X target frequency) was performed on the mean percentages of correct responses. A significant main effect of prime type was obtained [$F(1,43)=19.56$, $MS_e=0.0176$, $p < .001$]. (All results reported are $p < .01$ or beyond unless specifically stated otherwise). Targets primed with unrelated primes were identified more accurately than those targets presented with related primes. In all conditions, significant inhibition was obtained when targets were preceded by related primes. A significant main effect of target frequency was obtained [$F(1,43)=124.08$, $MS_e=0.0116$]. In both sparse and dense neighborhoods, high frequency targets were correctly identified significantly better than low frequency targets. A significant main effect of neighborhood density was also obtained [$F(1,43)=174.80$, $MS_e=0.0126$]. In all conditions, target words that came from sparse neighborhoods were identified more accurately than target words that came from dense neighborhoods. In addition to the main effects, the ANOVA revealed a significant interaction of neighborhood density X target frequency [$F(1,43)=21.68$, $MS_e=0.0087$]. The difference in response accuracy between sparse and dense neighborhoods was greater for low frequency targets (20.45% difference) than for high frequency targets (11.2% difference). These results suggest that the priming manipulation had a more robust effect when low frequency targets were used, compared to the conditions in which high frequency targets were used.

This experiment can be considered a successful replication of the main findings obtained in the Goldinger, Luce, and Pisoni (1989) experiment. The data show that priming with phonetically related primes produces inhibition of target word identification, compared to the identification of targets paired with unrelated primes. These results are consistent with three basic predictions of NAM: (1) high frequency targets should be identified more accurately than low frequency targets; (2) targets from sparse neighborhoods should be identified more accurately than targets from dense neighborhoods; and (3) phonetically related primes should inhibit the identification of target words. These predictions were based on the assumption

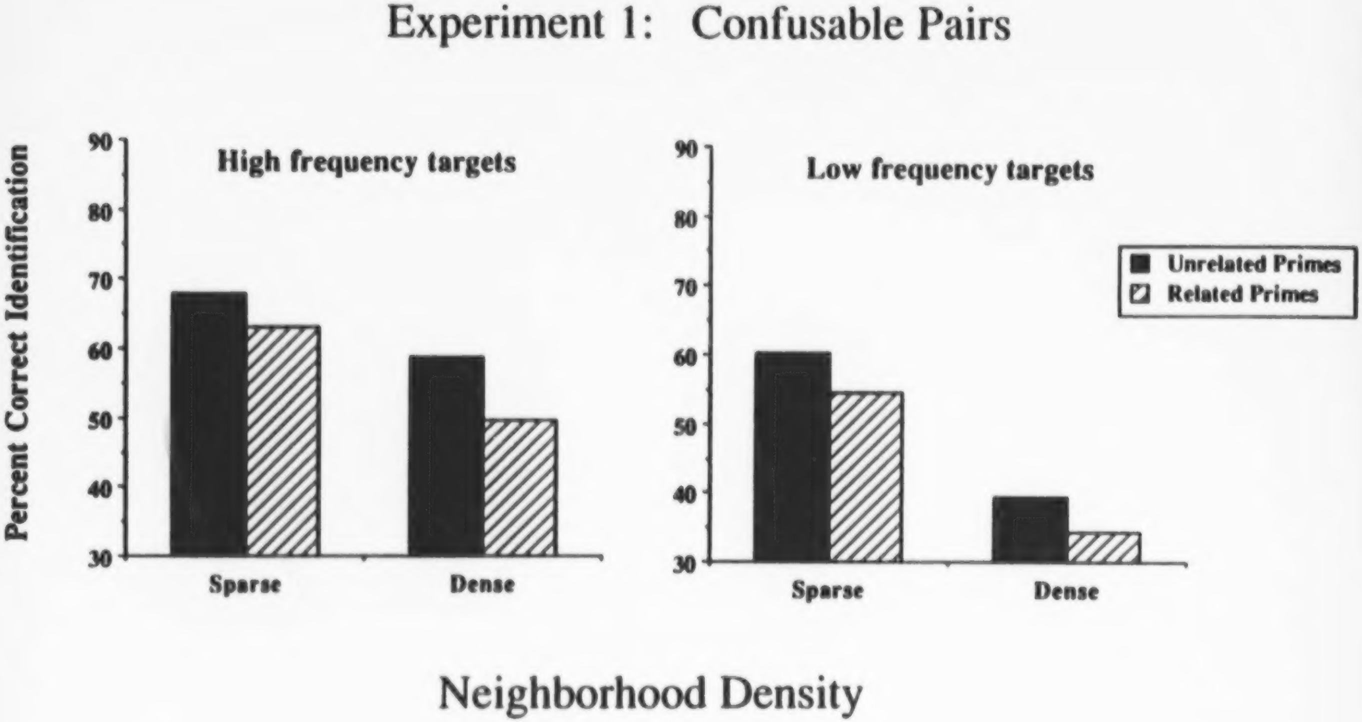


Figure 1. Percent correct identification for high and low frequency target words as a function of neighborhood density for related and unrelated primes. The light bars indicate conditions for primes that were related to their targets by phonetic similarity and dark bars indicate conditions with unrelated (neutral) primes. The mean percent correct identification for high frequency targets is shown on the left and performance for low frequency targets is shown on the right. Performance for sparse neighborhoods is indicated in the left half of each target frequency condition; performance for dense neighborhoods is indicated in the right half of each target frequency condition.

that residual activation remains in the target's neighborhood for some brief period after recognition of the prime.

In Experiment 1, we assumed (following Goldinger et al., (1989)) that the acoustic-phonetic similarity between the related primes and their targets was not explicit enough to be recognized by the subjects. Therefore, we predicted that subjects would not generate and/or apply any expectancies (or response strategies) based on this similarity. The only information we expected subjects to use to identify the target words was the acoustic-phonetic information provided by the target words themselves.

The data obtained in the post-task questionnaires supported this working assumption. When asked if there were any characteristics of the first word of each pair (the words in the clear) that the subject used to help identify or guess the second word of each pair (the words in noise), 61% of the subjects answered "yes" and 39% answered "no". However, of the 61% that answered "yes", only 37% said that the sounds of the first word (prime) helped them identify the second word (target), and only 40% of these subjects cited the initial sounds of the words as aids to identifying the target. So, overall, 0.9% of the 44 subjects said that the initial sounds of the prime helped them identify the target word. The remainder of the subjects who answered "yes" to this question said that they used relations between the words to help them identify the targets. The way the stimuli were selected, occasionally there could have been a semantic prime-target coupling, but overall the prime-target pairs had no obvious or predictable semantic relationships. Therefore, we assume that subjects who reported using semantic information to select their responses were most likely generating incorrect responses based on semantic expectations.

When subjects were asked if there were identical speech sounds produced in both words of the prime-target pairs, 77% said, incorrectly, that there were, and 23% said, that there were not. Of those that answered "yes" to this question, 30% said that the beginning sounds of the prime and target were identical. One subject stated, "The consonants seemed to be alike, but the vowels are more easily understood in the noise."

The stimuli were selected so that the prime and target were phonetically confusable with each other, but there were no identical phonemes present in any segments of the prime-target pairs. After reviewing the questionnaires, it appears that some subjects may have noticed the phonetic similarity between the primes and targets. However, since an overall effect of inhibition was obtained, we assume that subjects were not able to utilize this information in identifying the targets.

Finally, when asked if they had used a strategy, or thought a strategy was possible, 25% of the 44 subjects answered "yes" they had, and 75% answered "no" they had not. Of those subjects who said that they had used a strategy, none said that they used the initial sound of the prime to identify the target. Instead, they stated that they tried to use relations between the words, or tried to listen for similar vowel sounds. One subject responded, "No. For a while I started to believe that the second word began with the same letter as the first.

However, after a while I found this wasn't the case." Similarly, another subject stated, "I think it would take more trials to develop a strategy in order to find patterns." The majority of the subjects in this experiment said that they either tried to ignore the primes, or that they did not think any strategy was possible.

In summary, the results of this experiment show that even though many of the subjects may have detected the phonetic similarity between the related primes and their targets, they could not effectively utilize this information to aid them in their identification of the targets. Further, most of the subjects reported that they did not consciously use any type of strategy to help them in the experimental task. Thus, we assume that these subjects used only the acoustic-phonetic information provided to identify the targets. Finally, the overall effect of inhibition of target identification appears not to be due to a misuse of strategic processing, but rather appears to be due to the residual activation remaining in the similarity neighborhoods upon target presentation.

Experiment 1 was conducted, in part, to replicate the major findings of the experiment reported by Goldinger et al. (1989). Experiment 1 was conducted also to serve as a comparison condition to the method employed in Experiment 2. Whereas, in Experiment 1, the related prime-target pairs contained only similar, but non-identical, phonemes, the related prime-target pairs in Experiment 2 contain identical phonemes in word-initial position.

Experiment 2

Method

Subjects. Forty-four Indiana University undergraduates participated in partial fulfillment of requirements of an introductory psychology course. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing. None of the subjects had participated in Experiment 1.

Stimuli. The same targets and neutral primes that were used in Experiment 1 were used in Experiment 2. A new set of related primes were created. Unlike the related primes in Experiment 1, these related primes share one common initial phoneme with the targets.

As in Experiment 1, all new stimulus items satisfied several fundamental constraints: (1) All related primes were either two or three phonemes in length; (2) all words were monosyllabic; (3) all words were listed in the Kučera and Francis (1967) corpus; and (4) all words had a rated familiarity of 6.0 or above on a seven-point scale. The related primes were recorded in the same recording session as the stimuli described for Experiment 1.

Design. The experimental design and procedures employed in Experiment 2 were identical to those employed in Experiment 1.

Results and Discussion

The percentage of words correctly identified was determined for each subject. As in Experiment 1, only wholly correct identifications were scored as correct, although responses were corrected for simple spelling or typing errors prior to any statistical analyses. Figure 2 displays the results of the priming manipulation in Experiment 2. Light bars show performance for targets preceded by related primes, dark bars show performance for targets preceded by unrelated primes. Three main effects and one significant interaction are displayed in Figure 2: a main effect of prime type, a main effect of target frequency, a main effect of neighborhood density, and an interaction effect of target frequency and neighborhood density.

Insert Figure 2 about here

A three-way analysis of variance (prime type X neighborhood density X target frequency) was performed on the mean percentages of correct responses. A significant main effect of prime type was obtained [$F(1,43)=110.84$, $MS_e=0.0110$, $p < .001$]. (As in Experiment 1, all results reported are $p < .01$ or beyond, unless specifically stated otherwise). Across all conditions, significant facilitation was obtained when targets were preceded by related primes. A significant main effect of target frequency was obtained [$F(1,43)=140.91$, $MS_e=0.0104$]. In both sparse and dense neighborhoods, high frequency targets were correctly identified more accurately than low frequency targets. A significant main effect of neighborhood density was obtained [$F(1,43)=506.56$, $MS_e=0.0046$]. In all conditions, target words from sparse neighborhoods were identified more accurately than target words from dense neighborhoods. In addition to the main effects observed, the ANOVA revealed a significant interaction of neighborhhod density X target frequency [$F(1,43)=43.47$, $MS_e=0.0054$]. The difference in response accuracy between sparse and dense neighborhoods was greater for low frequency targets (21.54% difference) than for high frequency targets (11.2% difference). These results suggest that the priming manipulation had a more robust effect when low frequency targets were presented than when high frequency targets were presented.

The overall pattern of results clearly indicates that targets preceded by related primes that contained the same word-initial phoneme were identified more accurately than targets that were preceded by unrelated primes. In other words, facilitation of target identification was observed when the related prime shared a common word-initial phoneme with the target. These results replicate the previous results of Slowiaczek, et al. (1987) and are contrary to those observed in Experiment 1. The question of major interest in this experiment, however, is whether the phonological relationship between the related primes and their targets may have been explicit enough to be noticed and utilized strategically by the subjects. Therefore,

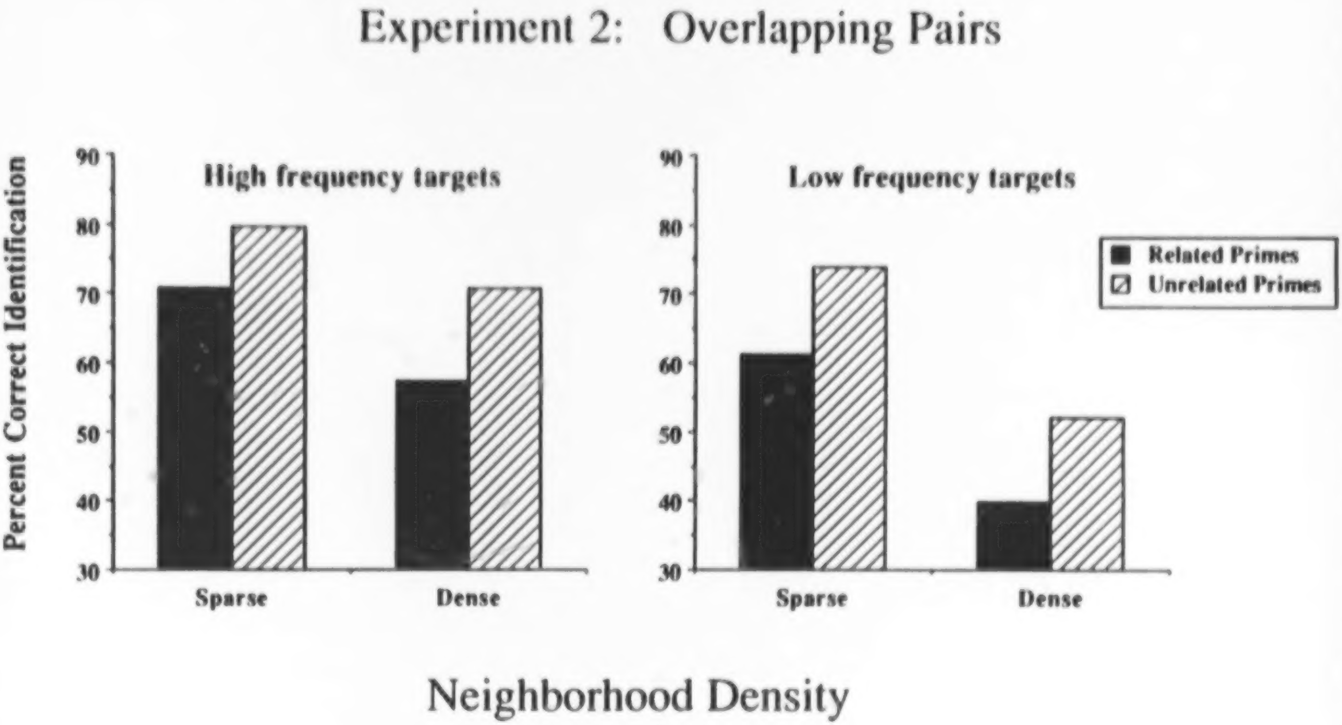


Figure 2. Percent correct identification for high and low frequency target words as a function of nieghborhood density. The light bars indicate conditions for primes that were related to their targets by initial phoneme overlap and dark bars indicate conditions with unrelated (neutral) primes and targets. The mean correct percent identification for high frequency targets is shown on the left and performance for low frequency targets is shown on the right. Performance for sparse neighborhoods is indicated in the left half of each target frequency condition; performance for dense neighborhoods is indicated in the right half of each target frequency condition.

it cannot confidently be stated that the results observed in Experiment 2 are due to a true priming effect. Instead, the subjects may have used the fact that there was a phonological overlap between the related primes and targets to limit their lexical search space and thereby overcome the residual neighborhood activation produced by low frequency primes. If this were the case, it could be considered a response strategy, as opposed to a "true" priming effect.

The data obtained in the post-task questionnaires supports the notion that subjects noticed the initial phonemic overlap between primes and targets, and that many subjects used this information strategically to help them identify the targets. When asked if there were any characteristics of the primes that helped identify the targets, 75% of the 44 subjects in the experiment responded "yes", and 25% responded "no". Of those who answered "yes", 79% said that the beginning sounds of the primes and targets were the same. One subject stated, "Sometimes the beginning sounds were the same and it helped me identify the second word [target]." It appears that many of the subjects in this experiment did notice the initial phoneme overlap between related primes and targets, and then utilized this information to aid in the identification of the target words.

When subjects were asked if identical speech sounds were present in both the primes and the targets, 84% reported that there were, and 16% reported that there were not. Of those who answered "yes", 61% said that these similarities occured in the beginnings of the prime-target pairs. Finally, the subjects were asked if they used any kind of response strategy to make the task easier for them, or if they believed any type of strategy was possible in this task. Of the 44 subjects, 37% said that they did use a strategy, and 63% said that they did not. Of the 37% that said they did use a strategy, 65% said that they tried to related the prime and the target, and 35% said that they used the initial sounds of the prime to help identify the target. It appears that even though the majority of the subjects in this experiment did say that they recognized the phonemic overlap between the primes and their targets, many of them allegedly did not to use this information to aid them in their identification of the targets. However, a large facilitation effect was obtained. One explanation may be that once the subjects realized that there was a phonemic overlap between the prime and target words in a pair, they *unconsciously* utilized this information in their identification of the targets.

General Discussion

Two priming effects were obtained in the present investigation. Both facilitation and inhibition of target word identification were observed in two experiments that differed only in the presence or absence of initial phonemic overlap between primes and targets. Subjects who were presented with phonetically related, non-overlapping prime-target pairs (e.g. *bull* and *veer*) showed inhibition of target word identification. Subjects who were presented with related prime-target pairs in which the word-initial phoneme was identical for both prime and target (e.g. *bull* and *beer*) showed facilitation of target word recognition. Neutral-target pairs that shared no phonetic similarity at all (e.g. *bull* and *gum*) were used as controls. All of the target words and all the neutral primes were identical in both experiments. Nevertheless, two opposite priming effects were still observed.

In Experiment 2, the subjects may have generated "expectancies" or biases toward certain responses. Subjects may have noticed the initial phoneme overlap in the related prime-target pairs. By using this information, subjects could generate a response strategy to help them identify the target words that were presented in noise. Thus, these subjects may have been responding on the basis of more limited sets of lexical candidates than those subjects in Experiment 1. If such a strategy were established, each trial would not be evaluated independently on the basis of acoustic-phonetic input. Instead, the facilitation of target word recognition could arise from a combination of acoustic information and subjects' expectancies. If this were the case, the facilitation observed in Experiment 2 could not be considered a true priming effect, but rather an effect of selective attention.

The questionnaires administered to subjects after the task lend support to the proposed hypotheses. The responses from the subjects in Experiment 1 indicated that, for most of the stimulus pairs, subjects did not notice any obvious relationship between the primes and their targets, and they stated that no information from the primes helped them identify the targets. One subject stated, "I didn't use any [strategy], and I really don't think any are possible. I saw no recognizable patterns to use." Many of the other subjects (75%) in Experiment 1 responded similarly. From both the patterns of priming observed and the questionnaire responses collected, it appears that the subjects in Experiment 1 were evaluating the targets mainly on the basis of their acoustic-phonetic properties and that higher level processes were not invoked to facilitate target identification.

Many of the subjects in Experiment 2, on the other hand, noticed and stated that the first word often began with the same "letter" as the second word, which "narrowed down the possibilities." Some subjects claimed that this information helped them identify the target word. As one subject noted, "I tried to listen to the second word with the first word's first consonant in mind." The overall effect of facilitation of target word identification cannot, however, be explained only in terms of a conscious bias on the part of the subject. Only 14% of the subjects in Experiment 2 reported consciously using the one phoneme overlap between the primes and targets of a pair to help them identify the targets. One subject states, "I tried to related the two words, but it wasn't always possible. I used no set strategy and I don't think the words were organized in any way that a strategy would be helpful." In fact, another subject states, "I didn't try any strategy really, because I didn't think there was any real pattern that could be picked up. I tried to avoid a strategy so I wouldn't be led by the sound of the first word."

In light of these comments, an alternate explanation of the results may be that some of the subjects in Experiment 2 *consciously* noticed the overlap between the prime and target words in a pair, but then *unconsciously* utilized this information when they identified the targets. This may also explain the fact that the error rates for the neutral prime-target pairs in both experiments were practically identical. Intuitively, if subjects in Experiment 2 were consciously biasing their decisions in a certain way, the percent correct identification of the targets from prime-target pairs that did not conform to this bias would be lower than if no strategies were being applied. Put another way, if subjects were led to believe that the target will begin with the same sound as the prime, they would have been more likely to make an error when this expectancy was not met. The data indicate, however, that in Experiment 1, where it was assumed that the subjects did not expect a certain kind of target to follow the prime, the error rates were nearly identical to those in Experiment 2, where an expectancy was assumed to operate.

Although there was no difference in overall error rates on unrelated prime-target trials, a subsequent error analysis revealed important differences between the responses given by the subjects in Experiments 1 and 2. Frequency counts were conducted on those errors in which the subjects' incorrect responses had the same initial phoneme as the preceding unrelated prime. Frequency counts were conducted on the total number of errors as well. In Experiment 1, in which the *related* priming trials employed only confusable between primes and targets, there were a total of 1801 errors committed on the *unrelated* priming trials, 160 of which showed "biased" responses, by the metric described above. In Experiment 2, in which the *related* priming trials employed overlapping phonemes between primes and targets, there were a total of 1333 errors committed on the *unrelated* priming trials, 238 of which showed "biased" responses. This is approximately a two-to-one ratio of biased guessing in Experiment 2 to biased guessing in Experiment 1, and the difference is statistically reliable [$\chi^2_1 = 41.93$, $p < .0001$]. This difference may be interpreted as evidence that the subjects in Experiment 2 did, in fact, learn the relevant relation between primes and targets during the course of the experiment, and that they utilized this information relatively frequently in response generation.

McLean and Shulman (1978) have suggested that expectancies are not a simple function of the direction of attention by the subject. Rather, attentional processes may be used to *construct* an expectancy, but this expectancy can then affect performance *independently* of the direction of attention. Therefore, in the context of Experiment 2, subjects could consciously notice the initial overlapping phoneme in the prime-target pairs, and then removed their attention from this characteristic of the experiment, but still maintain the expectancy of a phonemic overlap for the rest of the experiment. By McLean and Shulman's analysis, then, subjects may not think that a strategy was employed in the identification of the target where, in fact, responses were strategically selected.

Posner and Snyder (1975) propose a "pathway activation" explanation of attentional effects related to priming. Any item that is presented activates a pathway that contacts a memory representation of that item. This activation is automatic, requires no conscious attention, and will facilitate the processing of any subsequent item that shares the same pathway, but will not inhibit the processing of dissimilar items. If attention is conciously directed to the active information, however, inhibition of items that have dissimilar pathways will occur. The limited capacity of conscious attention will only facilitate the recognition of those items to which the attention is directed and unexpected items will be processed less efficiently. By Posner and Snyder's account, there is more than one way that attention can affect subjects' performance in a priming task.

In a later paper, Posner and Presti (1987) discuss the effects of attention on semantic priming. They state that semantic priming may occur automatically, even if the subject tries to ignore the prime. This automatic processing improves the processing of the primed item compared to an unprimed control item, but does not significantly inhibit the processing of items that are unrelated to the prime. When attention is directed toward the prime, however, there is a combination of facilitation of items that are related to the prime, and inhibition of items that are unrelated to the prime.

In a more recent paper, Farah (1989) proposed that qualitative differences exist between perceptual and semantic priming, and that the attentional mechanism that underlies each type of priming may be quite different. After examining the results of perceptual priming experiments concerning the direction of attention to spatial locations, Farah concludes that directing "...attention to a region in visual space increases subject's sensitivity for stimulus identification in that region" (p. 190). Nosofsky (1986; 1987) has made similar suggestions with regard to the role of selective attention in perceptual classification.

In several semantic priming experiments employing lexical decision tasks (Antos, 1979; Schvaneveld & McDonald, 1981), a misspelling detection task (O'Connor & Forster, 1981), and a sequential word-matching task (Johnston & Hale, 1984), it has been shown that semantic priming consists of large changes in bias and no changes of sensitivity. Thus, it seems as if two different kinds of attention are of importance in perceptual and semantic priming tasks. These differences in the utilization of attention may be the key to the differences observed in performance between the Slowiaczek, et al. (1987) and Goldinger, et al. (1989) experiments.

In summary, these experiments have shown that two very different priming effects can be obtained using nearly identical tasks. Experiment 1 yielded an overall effect of inhibition of target word identification, presumably due to residual activation in the targets' similarity neighborhoods from the earlier presentation of the primes. Experiment 2 yielded an overall effect of facilitation of target word identification. This appears to be largely an effect of selective attention on the part of the subjects. From these findings, we suggest that there are qualitatively different processes that underlie the inhibitory and facilitatiory effects observed in priming experiments. The above discussion of the facilitation and inhibition of target items due to attentional effects underscores the importance of further investigation of priming effects in spoken word recognition. Future studies in spoken word recogniton which draw conclusions from the priming paradigm must recognize and address the different effects of attention on subjects' performance.

References

Antos, S. J. (1979). Processing facilitation in a lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, **5**, 527-545.

Becker, C.A., & Killion, T.H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 389-401.

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, **82**, 407-428.

Farah, M. J. (1989). Semantic and perceptual priming: How similar are the underlying mechanisms? *Journal of Experimental Psychology: Human Perception and Performance*, **15**, 188-194.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Hillinger, M.L. (1980). Priming effects with phonemically similar words: The encoding-bias hypothesis reconsidered. *Memory & Cognition*, **8**, 115-123.

Jakimik, J., Cole, R.A., & Rudnicky, A.I. (1985). Sound and spelling in spoken word recognition. *Journal of Memory and Language*, **24**, 165-178.

Johnston, J. E., & Hale, B. L. (1984). The influence of prior context on word identification: Bias and sensitivity effects. In H. Bouma & D. Bouwhis (Eds.), *Attention and performance X* (pp. 243-255). Hillsdale, NJ: Erlbaum.

Kučera, F., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, R.I.: Brown University Press.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P.A., & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on speech perception progress report no. 7*. Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Luce, P.A., & Pisoni, D.B. (1987). Neighborhoods of words in the mental lexicon. Submitted.

Luce, R.D. (1959). *Individual choice behavior*. New York: Wiley.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continous speech. *Cognitive Psychology*, **10**, 29-63.

Martin, R. C., & Jensen, C. R. (1988). Phonological priming in the lexical decision task: A failure to replicate. *Memory & Cognition*, **16**, 505-521.

McClean, J. P., & Shulman, G. L. (1978). On the construction and maintenance of expectancies. *Quarterly Journal of Experimental Psychology*, **30**, 441-454.

Meyer, D.E., Schvaneveldt, R.W., & Ruddy, M.G. (1974). Functions of graphemic and phonemic codes in visual word recognition. *Memory & Cognition*, **2**, 309-321.

Neely, J.H. (1977). Semantic priming and retrieval from lexical memory: The roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, **106**, 226-254.

Neely, J.H., Schmidt, S.R., & Roediger, H.L. III (1983). Inhibition from related primes in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **9**, 196-211.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 1[illegible]57.

Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 87-108.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

O'Connor, R. E., & Forster, K. I. (1981). Criterion bias and search sequence bias in word recognition. *Memory & Cognition*, **9**, 78-92.

Posner, M. I. & Presti, D. E. (1987). Selective attention and cognitive control. *Trends in Neural Science*, **10**,13-17.

Posner, M. I., & Snyder, C. R. R. (1975). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt and S. Dornic (Eds.), *Attention and performance V.* (pp. 669-682). New York: Academic Press.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior*, **17**, 403-417.

Schvaneveldt, R. W., & McDonald, J. E. (1981). Semantic context and the encoding of words: Evidence for two modes of stimulus analysis. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 673-687.

Slowiaczek, L.M., Nusbaum, H.C., & Pisoni, D.B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**, 64-75.

Slowiaczek, L.M., & Pisoni, D.B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory & Cognition*, **14**, 230-237.

Tanenhaus, M.K., Flanigan, H.P., & Seidenberg, M.S. (1980). Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition*, **8**, 513-520.

Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, **26**, 608-631.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Appendix A

Word Identification Task Post-Test Questionnaire

Please take a few moments to carefully fill out this questionnaire. This questionnaire is a very important part of this experiment, and it would help us tremendously if you would answer each question as fully as possible.

****

1. What did the noise presented with the second word of each pair sound like to you?

----

2. Were there any characteristics of the first words of each pair (the words in the clear) that helped you identify or guess the second word of each pair (the words in noise)? (Circle one)

YES NO

If so, what were these characteristics?

----

3. Were the words presented in noise easily understandable?

----

4. Were the word pairs meaningful? That is, were there any obvious semantic relations between the two words of any pair (e.g. DOCTOR-NURSE...)? (Circle one)

YES NO

If there were semantic relations, did you notice a common theme to the experiment?

----

5. Were the words within any of the pairs "related" or "similar" to each other in any way? (Circle one)

YES NO

If so, how were they related, or similar to each other? Did this relationship appear to hold in all trials of the experiment, or only sometimes?

If there were no relations between the words within each pair, why do you suppose the words used were selected to be paired with each other?

----

6. In any of the word pairs that you heard, were there identical speech sounds produced in both words? (Circle one)

YES NO

If yes, in what part of the words did these similarities occur?

----

7. Did you consciously make this task easy for yourself in any way? Did you try to take advantage of any aspects of the words to help you identify the word in noise more easily? Did you use any type of strategy? If so, what strategy did you use? If you did not use any strategies, do you think any such strategies are possible?

Talker Variability and Spoken Word Recognition: A Developmental Study[1]

Brigette R. Oliver

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The author would like to thank David Pisoni and Linda B. Smith for their wisdom and guidance in the planning and completion of this study. The research reported here was supported, in part, by NIH Research Grant DC-0111-14.

Abstract

The effects of talker variability on spoken word recognition were studied developmentally in three, four, and five-year old children. Subjects listened to lists of words and identified each word by pointing to a picture in a six alternative visual display. The words and pictures were taken from the Word Intelligibility by Picture Identification test (WIPI). Three conditions were examined: single talker, single talker with varying amplitude, and multiple talker. Each child heard one list from a particular talker condition. Results showed a main effect of age and an interaction between talker condition and age. As expected, we found an increase in overall accuracy with age. However, the three-year olds did better in the single talker condition than the other two conditions (which did not differ). The four and five-year olds generally displayed ceiling levels of performance which may have obscured other differences. Further studies with younger children are underway to examine the nature of the differences in perception between single-talker and multi-talker word lists.

Talker Variability and Word Recognition: A Developmental Study

In the process of acquiring language, from the moment of birth forward, human listeners are regularly exposed to a wide variety of speakers who talk under an extremely diverse set of conditions. The acoustic characteristics of these voices vary quite dramatically from talker to talker, even when the same utterance is produced. Yet, this inherent variability apparently seems to cause little perceptual difficulty for the listener, even during the early years of childhood. Humans of all ages are amazingly adept at perceptually normalizing the acoustic cues that vary from talker to talker. This ability to adjust or compensate for talker variability is just one of the remarkable and as yet unsolved problems in speech perception. Normalization also occurs across differing phonetic, phonological, lexical and sentential contexts within a single speaker. In this report we will be concerned mainly with inter-talker variability and will ignore intra-talker normalization processes.

Talker normalization in normal conversational context *seems* so effortless that it is taken for granted most of the time. However, recent experimental evidence from our laboratory using both infants and adults demonstrates that there *are* costs incurred by this normalization process. In the area of adult word recognition, Mullenix, Pisoni and Martin (1989), replicated an earlier study conducted by Creelman (1957) who found that performance in a speech identification test was poorer for lists spoken by more than one talker than for lists spoken by only a single talker. Mullennix, et al. (1989) used *both* perceptual identification and naming paradigms and manipulated several variables such as signal to noise ratio, lexical density and word frequency. They found that talker variablity not only increased naming latency and reduced identification of degraded stimuli, but that these effects were more robust and less task dependent than structural differences related to word frequency and lexical density. The authors concluded that talker variability affects very early acoustic-phonetic processes and does not interact with higher level structural variables related to word recognition and lexical access.

Another study was carried out by Martin, Mullennix, Pisoni and Summers (1989) who investigated the effects of talker variability on *memory* for spoken lists of words. In a series of three experiments, Martin et al. (1989) manipulated talker variability, and several other variables, such as short-term memory load and retention interval using an interference task. They found significant effects for talker variability in the recall of early list items. Specifically, subjects in the multiple-talker condition recalled fewer list items from the primacy portion of the serial position curve than subjects in the single-talker condition. The results also showed better recall of visually presented preload digits in the single talker condition than in the multiple talker condition. The authors argued that their results were due to interference in the rehearsal process. Specifically, they claimed that recall of multiple talker items requires more processing resources than single talker items, thereby reducing the speed and/or efficiency of the rehearsal process used to transfer items into long-term memory.

Unfortunately, Martin et al. (1989) were unable to distinguish between two alternative hypotheses. One is that talker variabilility directly effects only early perceptual processes requiring extra resources for encoding. This effect would indirectly effect rehearsal capacity. A second possibility is that talker variability *directly* effects both early perceptual encoding *and* later rehearsal processes. In an effort to evaluate these two hypotheses and determine the exact locus of the effects, a study was conducted by Goldinger, Pisoni, and Logan (in review). Goldinger, et al manipulated talker condition, confusability of stimulus words, and rate of presentation in a serial recall paradigm. Rate of presentation was manipulated because it is believed to primarily effect rehearsal processes. The authors found that recall of multiple-talker lists was much more effected by the rate of presentation than was recall of single-talker lists and they, therefore, concluded that talker variability has an effect on *both* perceptual encoding and later rehearsal processes.

In summary, studies with adults investigating multiple talker effects have found costs associated with both initial perceptual processing and encoding of the signal into long-term memory. These costs have been shown to effect identification, naming, and recall of stimuli. Despite the fact that there are very few investigations into the effect of talker variability except with adults, there is some recent evidence that a cost of processing is evident as early as 2 months old.

Recent studies by Kuhl (1979, 1983) have shown that infants at six months of age who have learned to distinguish a vowel contrast in one voice can successfully generalize to different voices. This does not result from an inability to distinguish voices however, because newborns are capable of distinguishing their mothers voice from other voices. Based on these findings, it appears that infants have at least some rudimentary talker normalization processes. The next question is whether or not there are processing costs associated with normalization that can be measured in newborns. A recent study by Jusczyk, Pisoni and Mullenix (1989) attempted to address just these questions.

Using the high amplitude sucking (HAS) procedure, Jusczyk et al (1989) recently tested two-month old infants' perception of several speech contrasts. The critical test in this experiment was whether or not the infants would be able to detect a contrast between two syllables while ignoring changes in the speakers' voice(s). Results showed that infants were able to detect the change in all of the single talker conditions and in the multiple talker conditions where only the syllable changed. However, in the multiple talker condition, talker change group, where the talkers in the habituation phase and talkers in the test phases were different, infants could not discriminate the syllables. It is also notable that infants in multiple talker conditions took longer to habituate to the stimuli than infants in the single talker conditions. So, while infants are capable of talker normalization in various conditions, there appear to be processing costs for them as well as for adults.

Given this brief review of earlier findings, the motivation for the present experiment should be clear. The present investigation was an initial attempt to bridge the gap between the earlier infant and adult studies. Given the findings that infants are capable of normalizing, albeit at an apparent expense, we wanted to explore the developmental trend in a population just acquiring spoken language. More specifically, we wanted to investigate the effect that talker variability would have on word identification in three, four and five-year old children. In addition to using single and multiple talker conditions, we also included a third condition (single talker with varying amplitude) in order to assess the effects of another type of variability that was perceptually less complex than talker variability. We hoped that comparison of the results from the amplitude varying condition with the results from the other two conditions would shed more light on the nature of talker variability effects. In particular, if the earlier differences are due to perceptual normalization for a talker's voice, then the ampliitude varying condition should be similar to the single-talker condition. On the other hand, if young children respond differentially to any stimulus differences then both the amplitude varying and multiple-talker condition should be different from the control condition.

In this experiment we had several expectations. The simplest and most obvious prediction was an effect of age. We anticipated that older children would identify more words correctly than younger children across all three stimulus conditions. A more qualified prediction had to do with the effect of talker variability. Given that the pointing task was sensitive enough to measure a difference, we predicted better performance for the single talker than multiple talker lists with the varying amplitude condition falling somewhere in between. That the task would be sensitive enough to assess these differences was not at all a given. In word identification studies with adult listeners, the stimuli must be degraded with noise before an effect of talker variability becomes apparent. Based on the expectation of lower levels of performance with children in general, stimuli were presented in the clear, rather than embedded in noise. We view the present investigation as a pilot study and as much a test of the methodology as of the experimental hypotheses under consideration.

Method

Subjects. Forty-five children, 15 each at ages 3, 4 and 5, were recruited from the surrounding community by an ad in the local newspaper to participate in this experiment. The average for each age group was 3.69, 4.43 and 5.54 years. Five subjects at each age were randomly assigned to one of three talker conditions (single, amplitude varying and multiple). Each subject was run separately in a single session lasting approximately half an hour. Subjects were paid for their participation. Two subjects who partially completed the experiment received payment but were not included in the final analyses.

Stimulus materials. Three word lists (25 words each) from the Word Intelligibility by Picture Identification (WIPI) test were used as stimuli for this experiment (Ross & Lerman, 1970). The WIPI is a test designed to assess speech discrimination abilities of young children.

All words are monosyllabic and have an average familiarity of 6.957 (Nusbaub, Pisoni & Davis, 1984) and an average frequency of 99.45 (Kucera & Francis, 1967). In regular clinical applications, the person administering the test reads the test words aloud in a live voice. The child is shown a display of six pictures (a different display is used for each word) and is instructed to identify the picture of the word they hear. For our purposes, it was necessary to prerecord the lists of stimulus words on audio tapes and to play the tapes back to children over headphones. The rest of the procedure remained the same.

Seven male and seven female adults served as talkers to produce the stimulus materials. The 75 test words were presented randomly on a CRT screen in front of the talker who was seated in a sound-attenuated booth (IAC model 401A). Utterances were recorded on audiotape using an Electro-Voice model D054 microphone and an Ampex AG-500 tape recorder. The talkers were instructed to read the words aloud in a normal voice at a constant rate of speech. The words were then converted to digital form using a 12-bit analog-to-digital converter running at a 10-kHz sampling rate. The RMS amplitude levels of the words were digitally equated and the test words were edited using a digitally controlled waveform editor (Luce & Carrell, 1981). These operations resulted in a database of 75 words spoken by 14 talkers for a total of 1050 stimulus tokens.

The 1050 stimulus tokens were then presented to adult subjects to obtain identification scores. Six subjects participated in two, one-hour sessions. In one session, subjects heard the 525 tokens spoken by males, and in the other session they heard the 525 spoken by females. All stimuli were presented via headphones and subjects were instructed to record what word they heard by typing their response into a computer keyboard in front of them. Results were tallied and taken as a measure of the intelligibility of each token. The male talker with the highest identification score across all 75 tokens was chosen for use in the single-talker condition. All tokens from this talker were identified correctly by at least 86% (or 6 out of 7) of the judges. Audio tapes were made using this voice for each of the three different lists. For the varying amplitude condition, one third of the words from each list were randomly chosen to remain at the original amplitude. One third of the words were increased 3dB, and the remaining one-third were decreased 3 dB. The same male voice used in the single talker condition was also used to construct tapes for each of the three lists in the varying amplitude condition. The one male voice and two female voices with the lowest identification scores were eliminated from the data base leaving five male and five female voices from which to construct the multiple-talker tapes. Tokens were chosen at random with the requirement that each stimulus was identified at least above 86% (or 6 out of 7) correct identification. The number of words spoken by each of the 10 talkers was as equal as possible on each list (five of the voices spoke two words each, the other five voices spoke three words each). Lists were also balanced for gender of the speaker.

Design & Procedure. Testing occurred in a single session lasting approximately half an hour. All children were tested individually by the same experimenter in a small, well-lit room. All subjects were given a pure-tone screening test (at frequencies of 500, 1000, 2000, and 4000) prior to participation in the experiment to ensure that they had no major hearing problems. No children were rejected as a result of the pure-tone screening test. During the experiment, the child sat across from the experimenter either in the parent's lap or in a chair next to the parent. Parents were asked not to assist or coach the child in any way throughout the course of the experiment. Children were told that they were going to play a game with the experimenter where they could win a sticker. They were instructed to listen to the words presented over the earphones and point to the pictures of what they heard. A practice trial was completed to ensure that the child understood the instructions and could carry out the task.

In the practice trial, the child was shown a sample page with six pictures. The experimenter asked "What would you point to if you heard the word 'x'?" with 'x' being one of the 6 pictures (i.e.,'cat'). This procedure was repeated one or two more times to ensure that the child understood the nature of the task. None of the children had any difficulty understanding the experimental procedures. The experiment then began with stimulus words presented to the subject through TDH-39 headphones using a Uher 4000 Report-L tape recorder. The experimenter would say "show me this," or some analogous prompt, play a test word, then stop the tape recorder until the subject responded. Responses were recorded by the experimenter on a response sheet. The experiment continued in this fashion until the 25 words on the list were completed. Once or twice per list the experimenter would remind the child of the instructions. After the list was completed, the child got to choose a 'prize' sticker.

Results & Discussion

All analyses were carried out on arcsine transformations of the percent correct scores. Because the initial analyses indicated no effect of subject gender, the data were combined and will be presented as group means summed across males and females. Performance was extremely high at all ages. The overall mean was 20.79 or 83% of the words correct. Figure 1 shows the overall data for each of the three talker conditions.

Insert Figure 1 about here

Data were analysed in an ANOVA using the factors of age and talker condition. There were three age levels and three levels of talker condition: single-talker, varying-amplitude and multiple-talker. The ANOVA revealed a main effect of age [F(1,2)=7.40, p<.002] and a marginally significant interaction between age and talker condition [F(1,4)=2.35, p<.07]. As shown in Figure 2, accuracy increased with age as expected. Five-year olds did better than four-year olds who performed better than three-year olds. The overall mean percentages of of correct responses were 85.67, 84.53, and 72.53, respectively.

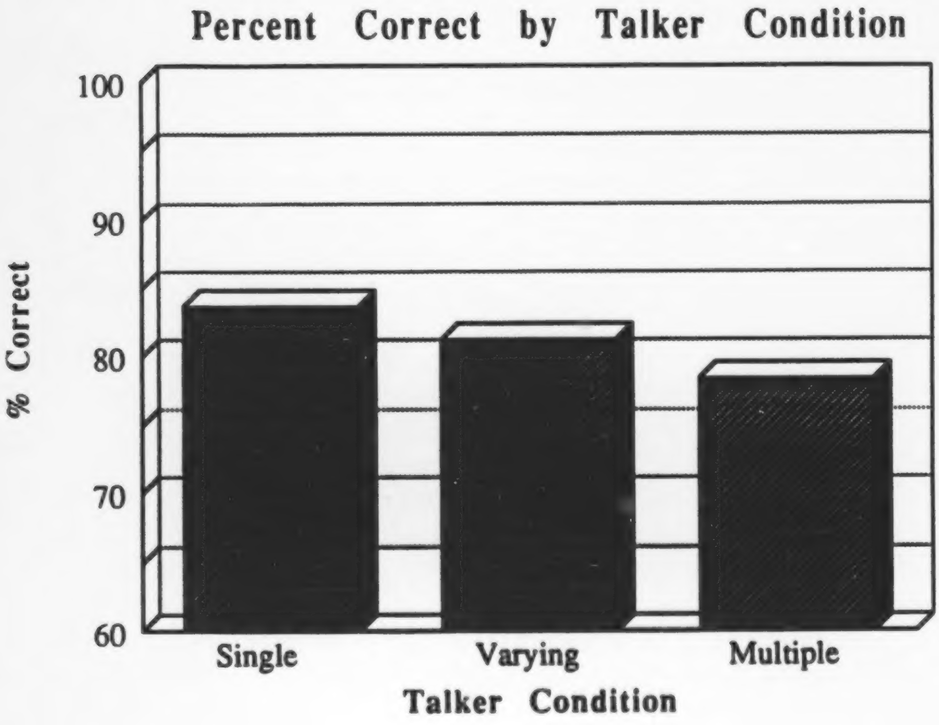


Figure 1. Average percent of words correct as a function of talker condition. Data is shown collapsed across ages.

Insert Figure 2 about here

When we further partition the data, it is easy to see the source of the interaction between age and talker condition. The relevant breakdown is shown in Table 1 and Figure 3. The three-year olds did well in the single-talker condition but much worse in the varying-amplitude and multiple-talker conditions. Four and Five-year olds did consistently well in all three talker conditions.

Insert Table 1 and Figure 3 about here

In an effort to further investigate the interaction, the data for each age group was analyzed separately. Analysis of variance on the three year olds revealed a main effect of talker condition [$F(1,2)=7.37$, $p<.008$]. Analysis of variance showed no significant effects for four or five-year olds. The significant results within the three-year old group appear to account for the marginal interaction across ages. Post-hoc Tukey's HSD analyses were conducted on the three-year olds' data to compare the results from the three talker conditions. These analyses showed that the single-talker condition differed significantly ($p< .05$) from the varying amplitude and multiple-talker conditions. The latter two conditions did not differ. From this result we can conclude that the single-talker list is easier than the varying-amplitude and multiple-talker lists for three year olds. We believe that large ceiling effects have resulted in no differences for the four and five year olds. A more difficult task that would result in lower performance is probably needed to elicit any differences that may be measurable for the older children.

Further analysis of the data remains to be conducted. It is possible that an item analysis would shed more light on some of the findings. For instance, if it were true that children missed more in the varying-amplitude list due to the one third of the words that were lower in amplitude, then we would have an explanation for the poor performance in that condition.

In summary, the results of the present study further emphasize that the right task is critical for measuring the effects of talker variability in young children. While the data from the three-year olds show significant differences between conditions, we believe ceiling effects in the four and five-year olds' data prevent the possibility of observing any significant differences in those groups. Several next steps are possible, many of them being the same steps often taken in studies with adults. First, noise could be added to the tapes in order to make the task more difficult. Second, we could add a time limit in order to pressure the children to respond quickly and decrease ceiling effects. This manipulation would probably be difficult or impossible with children this young. Third, instead of imposing a time limit,

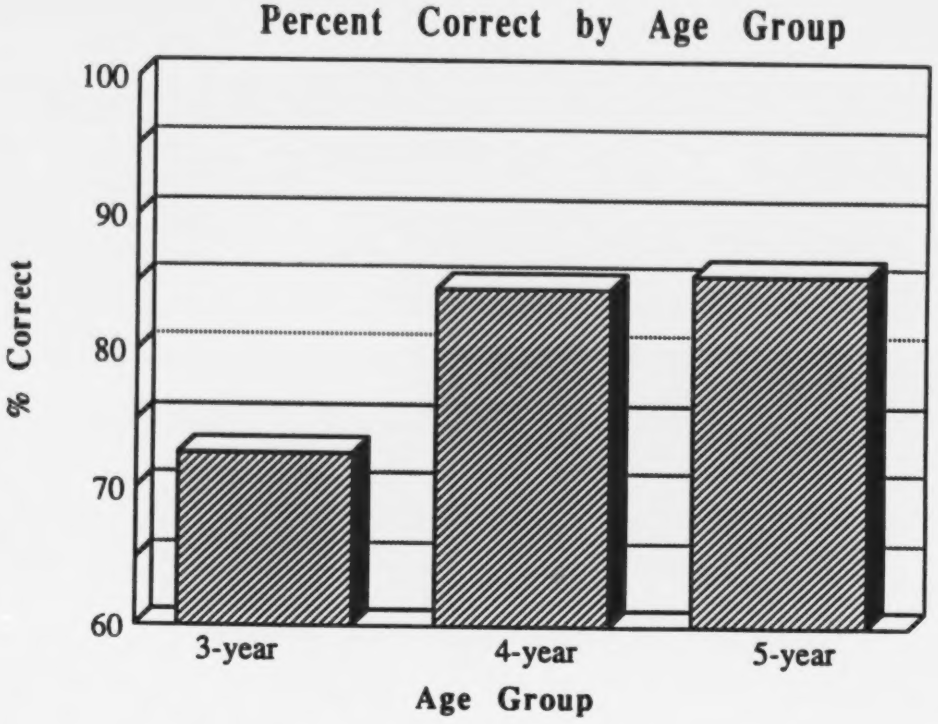


Figure 2. Average percent of words correct as a function of age. Data is shown collapsed across talker condition.

Table 1

Mean percent correct by age and talker condition.

Age (yrs)	Talker condition		
	Single	Variable	Multiple
Three	82.4	68.0	67.2
Four	81.6	91.2	80.8
Five	86.4	84.0	87.2

*N=5 for all cells.

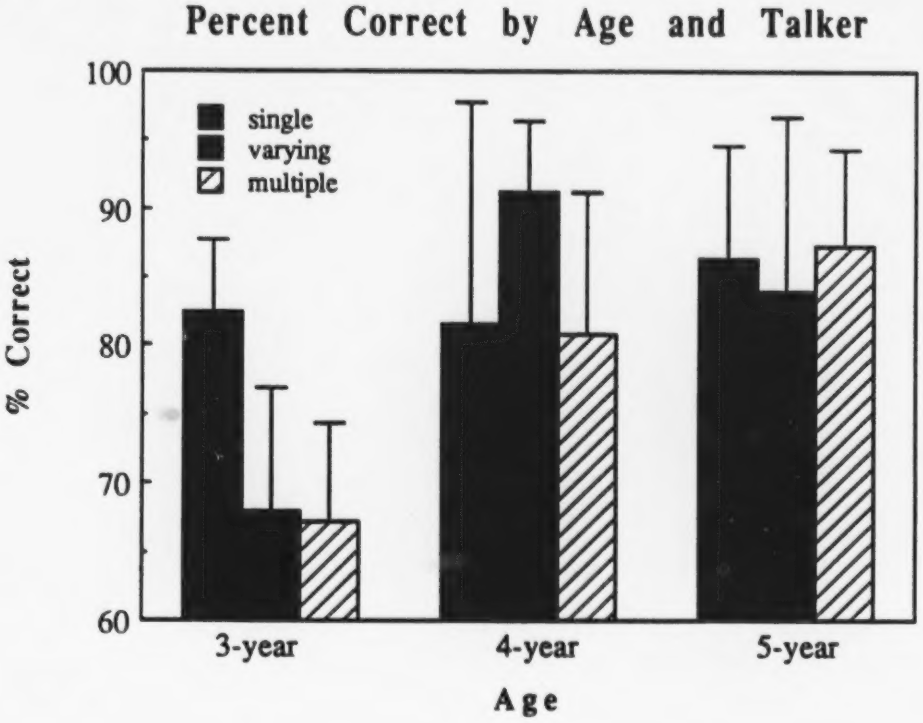


Figure 3. Average percent of words correct shown as a function of age and talker condition.

a more feasible option is to measure reaction times in a naming task and look for differences there. A fourth possibility is to move down to two-year olds whose level of function should be less than the older children, thereby avoiding the possibilty of ceiling effects. This is the option we are currently pursuing in our next study.

References

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustic Society of America*, **29**, 655.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (in press). On the nature of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. W. (1989). Effects of talker variability on speech perception by 2-month-old infants. *Research on speech perception progress report no. 15*. Bloomingtion IN: Speech Research Laboratory, Psychology Department, Indiana University.

Kucera, F., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, **66**, 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, **6**, 263–285.

Luce, P. A., & Carrell, T. D. (1981). Creating and editing waveforms using WAVES. Research on speech perception progress report no. 7. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of Talker Variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **15**, 676–684.

Mullennix, J. W., Pisoni, D. B., & Martin, D. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365–378.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Ross, M. & Lerman, J. (1970). A picture identification test for hearing impaired children. *Journal of Speech and Hearing Research*, **13**, 44-53.

Effects of Cognitive Workload on Speech Production: Acoustic Analyses[1]

W. Van Summers[2], David B. Pisoni, Robert H. Bernacki

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This work was supported by a contract from Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, Contract No. AF-F-33615-86-C-0549 to Indiana University, Bloomington, Indiana.

[2] Now at Army Audiology and Speech Center, Walter Reed Army Center, Washington, DC 20307.

Abstract

The work presented here examined the effects of cognitive workload on speech production. Cognitive workload was manipulated by having subjects perform a visual compensatory tracking task while they were speaking test sentences. Sentences produced in this workload condition were compared with control sentences which were produced in a no workload condition. Subjects produced utterances in the workload condition with increased amplitude, increased amplitude variability, decreased spectral tilt, increased F0 variability, and increased speaking rate. These changes involve both laryngeal and sublaryngeal structures and changes in articulatory timing. There was no evidence of vowel reduction or other changes in subjects' abilities to acheive supralaryngeal articulatory targets.

Effects of Cognitive Workload on Speech Production: Acoustic Analyses

Attentional and cognitive demands placed upon pilots, flight controllers, and others involved in information-intensive jobs may influence the acoustic characteristics of their speech in demanding situations. Very little research has explored whether consistent changes can be identified in the characteristics of utterances produced in demanding or "high-workload" environments. This type of research could have several important applications. First, if speech characteristics could be identified which correlate with the level of workload an operator is experiencing, this information could be used in training and selecting operators or in testing environments for their human-factors acceptability. This information could also be important in the design of speech-recognition devices which may operate in high-workload settings. These devices must be able to tolerate changes in acoustic characteristics that occur as a result of variability in workload. The present study was aimed at exploring whether consistent changes in speech could be identified which were the result of changes in the attentional and cognitive demands of the environment.

Previous research involving workload tasks have generally assumed that workload increases are associated with increased psychological stress (e.g., Hecker, Stevens, von Bismark and Williams, 1968; Tolkmitt and Scherer, 1986). Therefore, the results of these studies have often been equated with studies in which stress is manipulated through exposure to aversive stimulation, instructions requiring subjects to lie to the experimenter (or an accomplice) or other means of increasing emotional stress (Scherer, 1979). In addition, the majority of previous studies concerned with these issues have focused on fundamental frequency (F0) characteristics as an indicator of workload or stress (see, for example, Williams and Stevens, 1969; Kuroda, Fujiwara, Okamura and Utsuki, 1976; Tolkmitt and Scherer, 1986). Few studies have examined a larger array of acoustic characteristics in an effort to produce a more complete description of the effects of increased workload on the speech signal (but see Hansen, 1988). The present study examined how changes in cognitive workload affect various acoustic characteristics of the speech signal and whether changes that can be associated with increased workload are similar to changes that have previously been ascribed to increased emotional stress.

In the present study, cognitive workload was manipulated by requiring speakers to perform an attention-demanding secondary task while speaking. The task chosen was a compensatory tracking task which was first described in Jex, McDonnell and Phatak (1966). The tracking task will be referred to as the "JEX" task hereafter. The task involved manipulating a joystick in order to keep a pointer centered between two boundaries on a computer screen (see Figure 1). The program deflected the pointer away from the center position and the subject was required to continuously compensate for the movement of the pointer by manipulating the joystick in order to keep it from crashing into one of the boundaries. Phrases which the subjects were required to produce were visually presented on the computer screen while the subjects continued to perform the JEX task.

Insert Figure 1 about here

Method

Subjects. Five male native English speakers were recruited as subjects. Three subjects (ME, TG, and EG) were psychology graduates student who were paid for their participation. Two subjects (MC and SL) were members of the laboratory and participated as part of their routine duties. All speakers were naive to the purpose of the study. None of the speakers reported a hearing or speech problem at the time of testing.

Procedure. Subjects were run individually in a single-walled sound-attenuated booth (IAC Model 401A). The subject was seated comfortably facing a video screen which contained the JEX task display and (during sessions in which speech was collected) the phrases to be produced. The subject wore a headset fitted with an Electrovoice condenser microphone (Model C090) which was attached to the headset with an adjustable boom. Once adjusted, the microphone remained at a fixed distance of 4 inches from the subject's lips throughout the experiment. Subjects wore the headset during training sessions on the JEX task and during the experimental sessions in which utterances were collected.

Subjects were trained on the JEX task alone for several days. When a subject was able to consistently perform the task at a fairly high level of difficulty, the actual experiment was conducted. During the experiment, subjects simultaneously performed the JEX task and produced phrases presented on the video screen. Test phrases consisted of /h/-vowel-/d/ utterances in the sentence frame: "Say hVd again". The English vowels, /i,ɩ,ɑ,æ,ɚ,ɷ,u,ʌ, o^{ɷ},e/ appeared in the hVd context. During the actual experiment, JEX task difficulty was set to 80% of the level attained by the subject during training. Test phrases were only presented while the subject was performing the JEX task at this level of difficulty.

For each subject, utterances were collected in two experimental sessions. Each session consisted of 4 blocks of 20 trials with each of the 10 phrases presented twice within each block. Blocks of trials alternated between "JEX" and "control" blocks. During JEX blocks, subjects performed the JEX task while producing the test phrases; during control blocks, subjects produced the test utterances without performing any simultaneous task. For each subject, a total of 8 tokens of each phrase were produced in each condition [1]

---

[1] For subject EG, we report data from experimental session 2 only. The level of JEX task difficulty (80% of the level attained during training) used in experimental session 1 was apparently too low for this subject. He performed the task without any crashes for the entire session and did not appear to be under any workload. JEX task difficulty was increased in session 2. With this increase in difficulty, performance on the JEX task was similar to the performance of the other subjects. The exclusion of session 1 data for EG left 4 tokens of each utterance available from each condition.

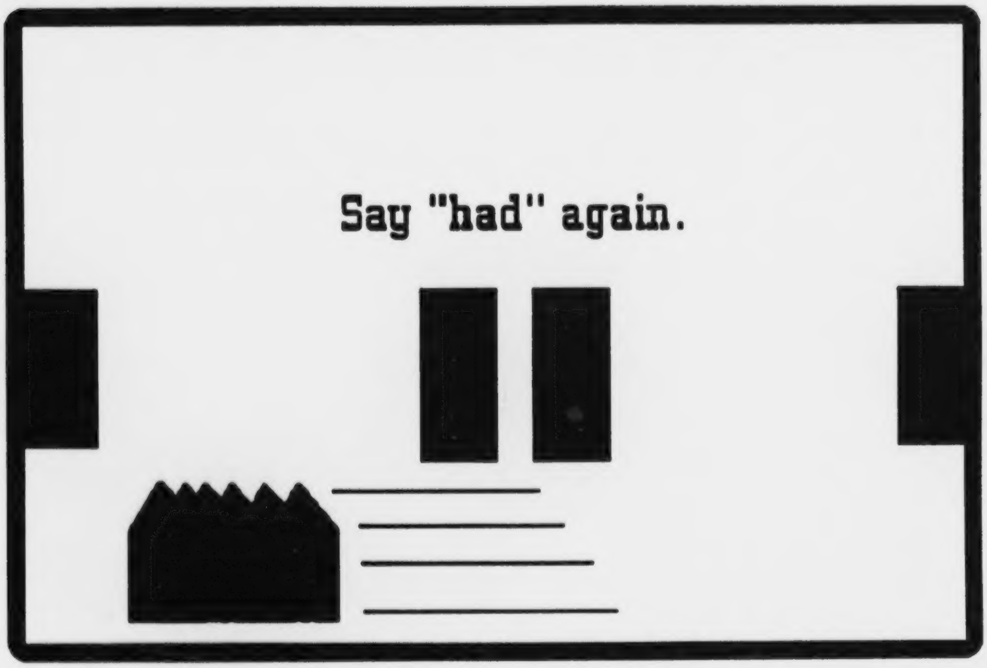


Figure 1. Illustration of the JEX compensatory tracking procedure. The subjects' task is to keep the moving pointer located at the bottom of the display from crashing into the sides of the display. During speech-collection intervals, sentences were presented in the top portion of the display as shown.

Speech Signal Processing. Test utterances were analyzed using digital signal processing techniques. Utterances were sampled and digitized on-line by a VAX 11/750 computer during the experimental sessions. Utterances were first low-pass filtered at 4.8 kHz and then sampled at a rate of 10 kHz using a 16 bit A/D converter (Digital Sound Corporation Model 2000). Each utterance was sampled into a separate waveform file.

Linear predictive coding (LPC) analysis was performed on each waveform file. LPC coefficients were calculated every 12.8 ms using the autocorrelation method with a 25.6 ms Hamming window. Fourteen linear prediction coefficients were used in the LPC analyses. The LPC coefficients were then used to calculate the short-term spectrum and overall power level of each analysis frame (window). Formant frequencies, bandwidths, and amplitudes were also calculated for each frame from the LPC coefficients. In addition, a pitch extraction algorithm was employed to determine if a given frame was voiced or voiceless and, for voiced frames, to estimate the fundamental frequency (F0).

Total duration for each phrase was determined by visual inspection and measurement from a CRT display which simultaneously presented the waveform along with time-aligned, frame-by-frame plots of amplitude, F0 (for voiced frames), and formant parameters. Cursor controls were used to locate the onset and offset of each utterance. The onset and offset of the /h/ frication, vowel, and /d/ closure segments from the hVd portion of each utterance were also identified and labelled. Following identification of utterance and segment boundaries, a program stored durational and RMS energy information for each utterance and segment. Fundamental frequency and formant frequency information were also stored for the phrase and for the vowel of the hVd portion of the phrase.

Results and Discussion

The influence of cognitive workload on various acoustic characteristics of the test utterances is described below. In each case, an analysis of variance was used to determine whether workload had a significant effect on a given acoustic measure. Separate analyses were carried out for each speaker. The analyses used phrase and workload condition (JEX or control) as independent variables and a p value of .05 as the critical value in all tests of statistical significance. The presentation of results will focus on the effect of workload on the various acoustic measures. The phrase variable will be discussed only in cases where a significant phrase X workload interaction was observed.

Amplitude. The upper panel of Figure 2 shows amplitudes averaged across entire phrases for utterances from the JEX and control conditions. The data are plotted separately by speaker. The lower panel of the figure shows amplitudes at the segmental level. Amplitudes of the /h/ frication, vowel, and /d/ closure portion of each hVd utterance are shown for each workload condition. An asterisk above a pair of bars indicates a significant difference between values in the JEX and control conditions for a particular speaker.

Insert Figure 2 about here

As the figure shows, there was a tendency for amplitude to increase in the JEX condition. The pattern is very consistent for speakers SL, ME, and TG who showed significantly higher amplitudes in the JEX condition for the entire phrase and for the separate /h/, vowel, and /d/ closure segments. Speaker MC showed significantly higher amplitudes in the JEX condition for the vowel and /d/ closure segments. Speaker EG did not show as clear a pattern of amplitude increases under workload as the other speakers. For this speaker, /h/ frication amplitude was significantly higher in the control condition than in the JEX condition. However, this significant main effect was mediated by a significant phrase X workload condition interaction. For EG, /h/ frication amplitude was higher in the control condition in 7 of the 10 vowel contexts. Of all of the analyses reported in this study, this was the only case of a significant phrase X workload interaction.

Amplitude variability (across utterances). Along with an increase in mean amplitude, amplitude variability from one utterance to the next also tended to increase in the workload condition. Figure 3 shows standard deviations of phrase amplitude across utterances for each condition. For the entire phrase, four of the five subjects showed an increase in amplitude variability in the JEX condition. For three of these subjects, the increase in variability was statistically significant. One subject showed the opposite pattern with significantly less amplitude variability in the JEX condition. As the lower panel of the figure shows, the pattern just described for the entire phrase was also true for amplitude variability of vowels and /d/ closure segments in the h-vowel-d context. In each case, four of the five subjects showed greater amplitude variability when performing the workload task. The effect was statistically significant for three of these four speakers in the case of vowels but was only signficant for one speaker in the case of /d/ closure.

Insert Figure 3 about here

Spectral tilt. Amplitude increases are often correlated with changes in "spectral tilt". That is, high amplitude utterances generally show flatter spectra with relatively more high frequency energy than is seen in low amplitude utterances. We examined the long-term spectra of the hVd vowels in each condition to determine if the amplitude increases seen in the JEX task were correlated with decreased spectral tilt. Figure 4 shows the difference in energy between JEX-condition and control-condition vowels across 40-Hz linear frequency bands.

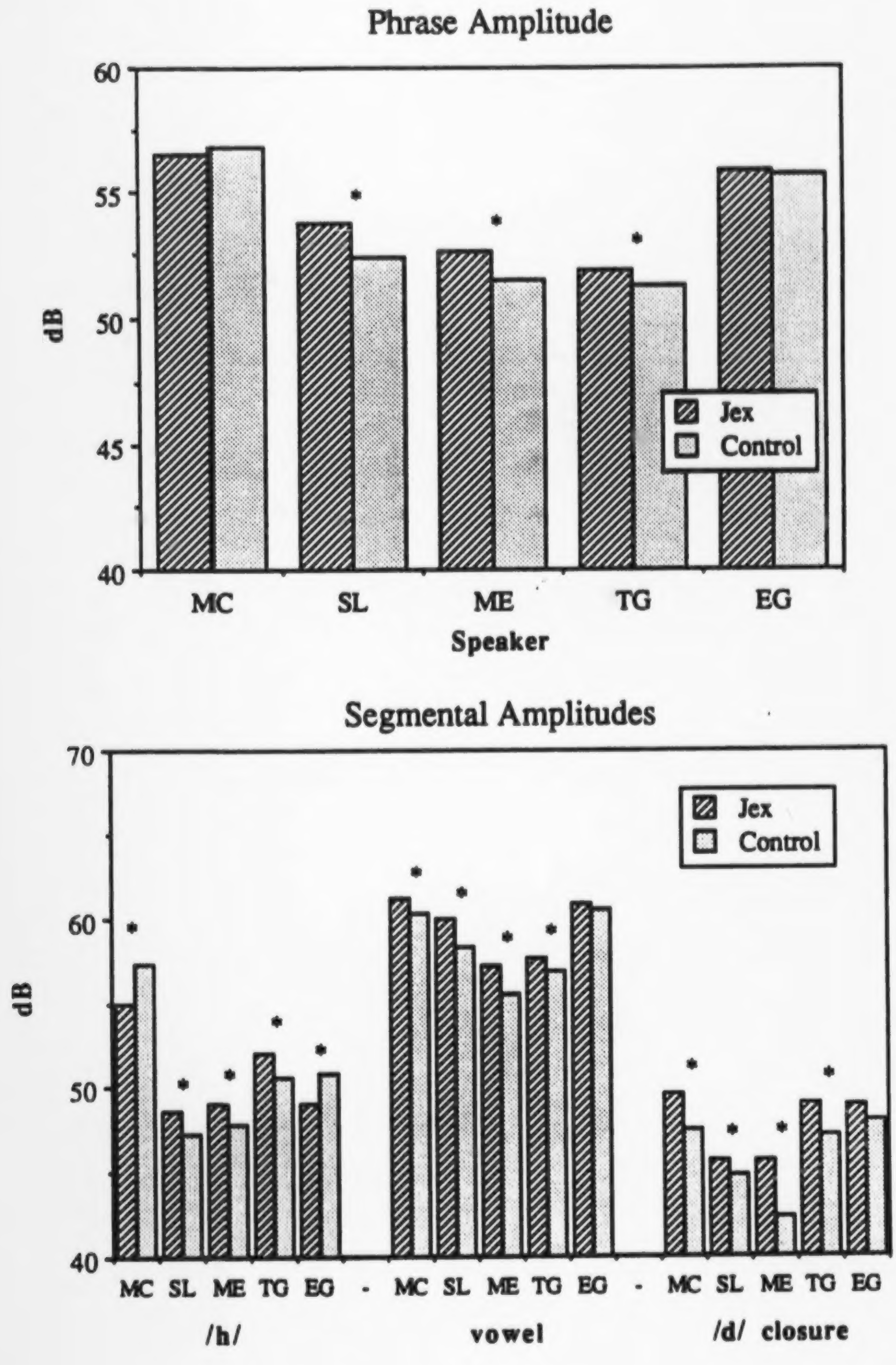
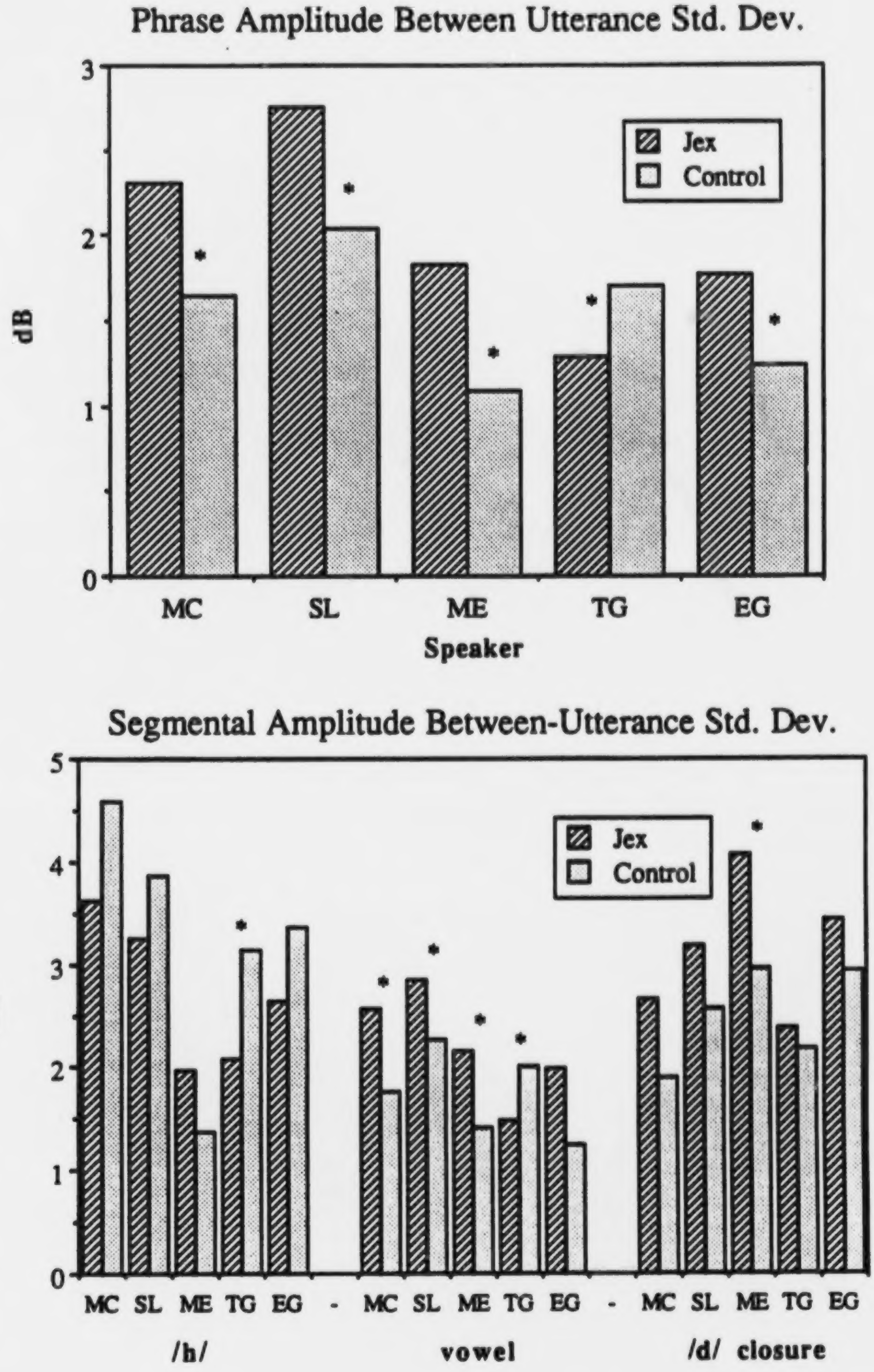


Figure 2. Phrase amplitude (upper pannel) and segmental amplitudes for "Say hVd again" utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.


Phrase Amplitude Between Utterance Std. Dev.

Segmental Amplitude Between-Utterance Std. Dev.

Figure 3. Between-utterance standard deviations for phrase amplitudes (upper panel) and segmental amplitudes. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

Insert Figure 4 about here

A positive slope in these figures indicates that the difference in energy between JEX-condition and control-condition vowels is increasing with frequency. This is the pattern seen for speakers MC, SL, ME, and EG. Thus, as in the amplitude data, four of the five subjects show a consistent pattern. However, the four subjects who showed changes in spectral tilt across conditions are not the same four who showed amplitude differences. Subject EG, the subject who did not show significant amplitude differences across workload conditions, shows one of the clearest cases of changes in spectral tilt. Overall, the data suggest that the workload task produces effects on spectral tilt that are not always correlated with changes in overall amplitude.

Hansen (1988) has provided further evidence that decreases in spectral tilt under workload are not necessarily linked to amplitude increases. For the both the JEX task and the dual task examined by Hansen, spectral tilt decreased under workload without any amplitude increase.

Fundamental frequency. We also analyzed fundamental frequency for each phrase and for the hVd vowel segments in each condition. Figure 5 shows mean F0 values for the phrase and hVd vowels in each condition for each talker. Two speakers (MC and SL) showed a significant increase in F0 for the entire phrase and for the hVd vowel when performing the JEX task. The pattern of F0 increase under workload was not replicated for either the phrase or hVd vowel in the other three subjects' data.

Insert Figure 5 about here

The absence of a consistent effect of workload on mean F0 values was also reported by Hansen (1988). Hansen examined speech produced while performing the JEX task and a "dual task" requiring the simultaneous performance of two tracking tasks while speaking. Neither of these workload tasks had any consistent effect on mean F0. Conversely, previous research examining the effects of emotional stress on speech have generally reported increases in mean F0 accompanying increased stress (Williams and Stevens, 1969; Kuroda et al., 1976; Streeter, MacDonald, Apple, Krause and Galotti, 1983).

Fundamental frequency variability (within utterances). A different characteristic of the F0 data does, however, show a fairly consistent pattern across workload conditions. Figure 6 shows the standard deviations of the frame-by-frame F0 values for each phrase and for each hVd vowel in each condition. As the figure shows, three subjects showed a significant

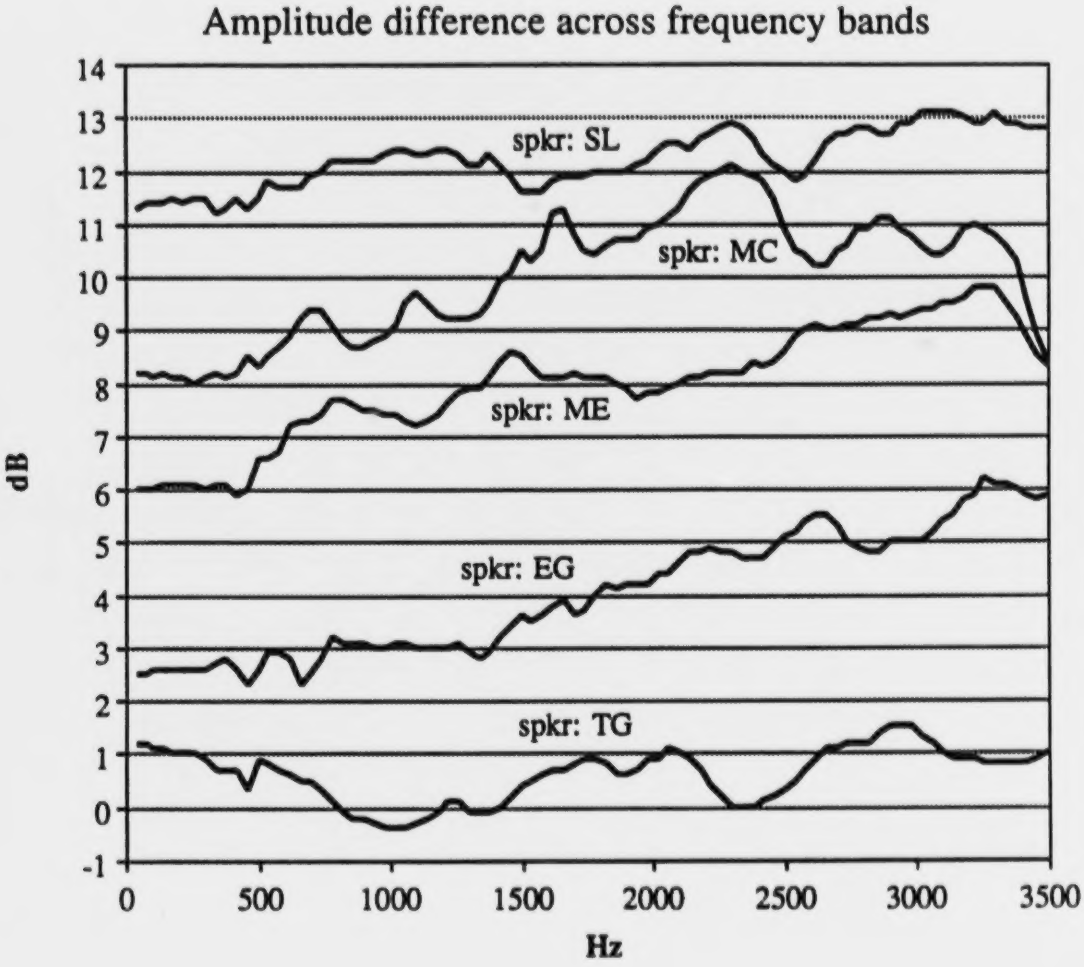


Figure 4. Mean difference in energy between utterances produced in the JEX and control conditions across frequency bands. Values are collapsed across utterances and presented separately for each speaker. For clarity of presentation, the traces for speakers EG, ME, MC and SL have been elevated by 2.5, 5, 7.5 and 10 dB respectively.

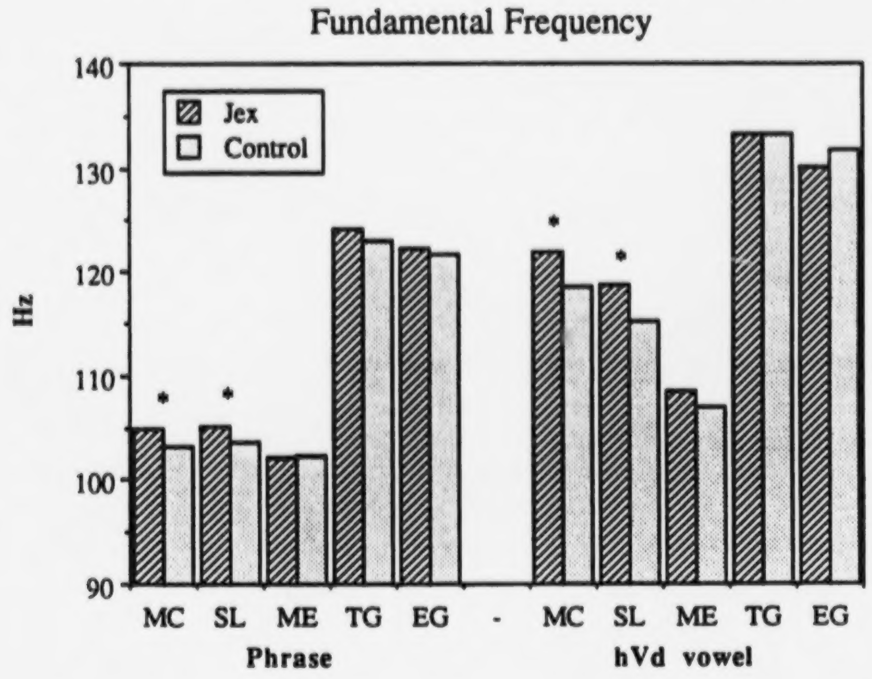


Figure 5. Mean fundamental frequency values for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

reduction in F0 variability when performing the JEX task. A fourth subject showed the same general pattern, although the difference was not significant. The pattern is not seen in the vowel data. Given that F0 variability decreases for the entire phrase but not the vowel, the pattern is more likely due to a flattening of the overall F0 contour rather than a decrease in period-to-period F0 variability or "vocal jitter." In other words, the whole phrase is apparently produced using a monotone pitch when the subject is under workload.

Insert Figure 6 about here

The decrease in F0 variability under workload is interesting in light of previous research examining the performance of the Psychological Stress Evaluator (PSE), a commercially-available "vocal lie detector". The PSE responds to an 8-14 Hz frequency modulation (FM) in the vocal signal. Brenner, Branscomb and Schwarz (1979) report that this frequency modulation is decreased when subjects are required to perform a speeded arithmetic task. This type of decrease in F0 modulation could produce the decrease in F0 variability reported above although it should be observed for individual vowels even more readily than for entire phrases.

Duration. Figure 7 shows the effect of cognitive workload on phrase durations and segmental durations. Four of the five speakers showed significantly shorter overall phrase durations while performing the JEX task. One speaker showed the opposite pattern with longer phrase durations while under workload. This speaker showed the smallest change in phrase duration across conditions. Segmental durations also tended to be reduced while performing the JEX task. The four speakers who showed shorter phrase durations in the JEX condition also tended to show shorter /h/ frication durations and shorter /d/ closure durations. Vowel duration (in hVd words) was less consistently affected by the workload condition. The durational shortening observed for the entire phrase, the /h/ frication, and the /d/ closure, replicate results mentioned briefly in Hecker et al. (1968).

Insert Figure 7 about here

Given that the vowel in the hVd contexts was the only part of the phrase containing "new" information from trial-to-trial, speakers may have treated the production of this vowel as more important than the production of surrounding context. This may explain why vowel duration was not consistently reduced in the JEX condition while other segmental durations were reduced in the remainder of the utterance.

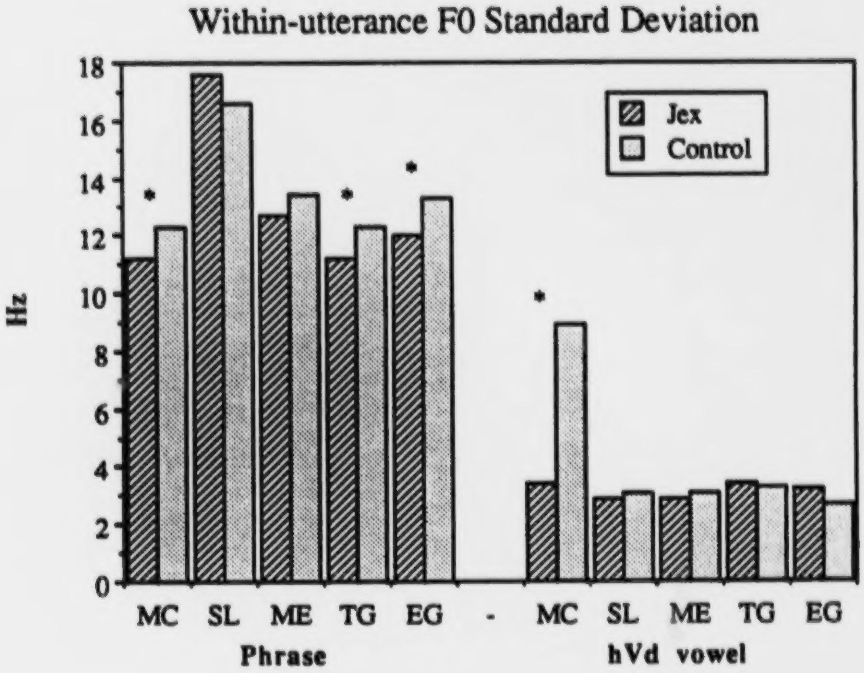
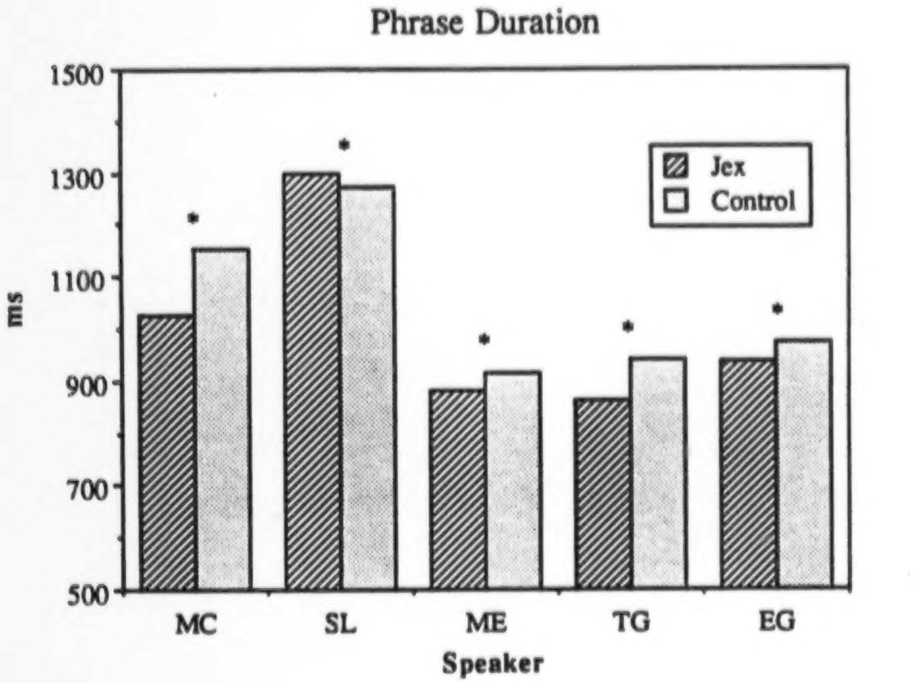


Figure 6. Mean within-utterance F0 standard deviations for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.



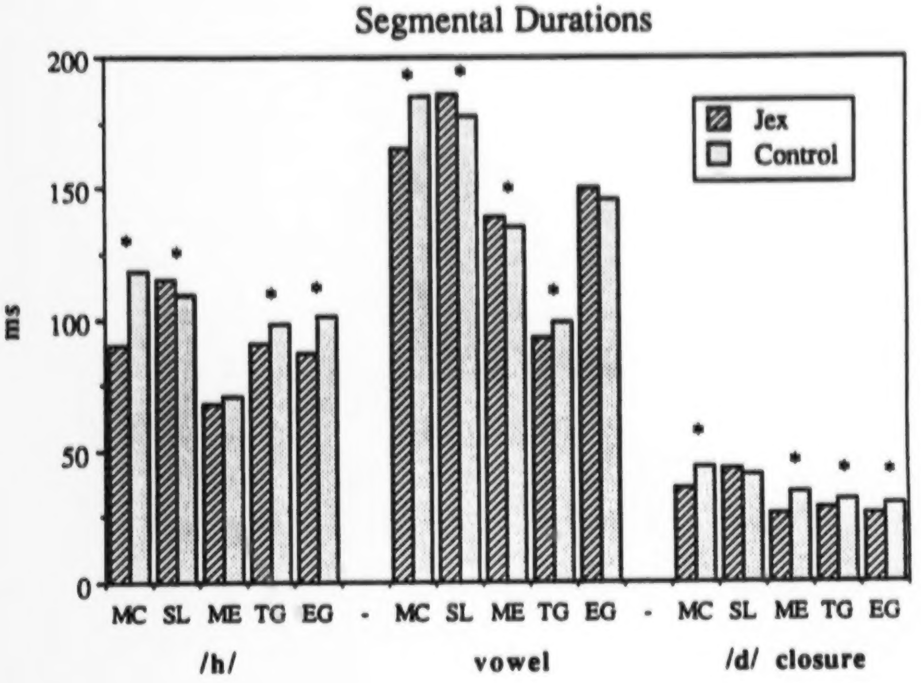


Figure 7. Mean phrase duration (upper pannel) and mean segmental duration values for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

Goldman-Eisler (1968) reports data on speech pauses and spontaneity that may be related the present findings. Her data suggest that hesitations decrease and fluency increases as subjects repeat a given passage. This may explain why subjects increase their speech rate (i.e., decreased durations) for the part of the phrase which they continued to repeat. Presumably this decrease in duration would allow more time to be allocated to the workload task. On the other hand, the identity of the hVd vowel changed from trial-to-trial so that this increase in fluency/rate did not occur for this portion of the phrase.

Formant frequencies. Workload did not have a clear influence on the frequencies or bandwidths of the first three formants for any of the five speakers. Thus, it appears that workload had a greater influence on sub-laryngeal and laryngeal (source-related) functions and speech timing than it did on the supralaryngeal control of speech.

Summary and Conclusions

Very little previous research has attempted to identify consistent changes that occur in the acoustic-phonetic properties of speech produced in severe environments. Research in this area may have important implications for human-to-human and human-to-machine speech communication in demanding environments such as cockpits and air traffic control towers.

The present results show that increased cognitive workload produces a number of effects on the acoustic-phonetic properties of speech. Utterances produced under cognitive workload show higher amplitudes and greater amplitude variability between utterances. Spectral tilt was reduced for vowels produced under workload and this change in tilt was not always correlated with a change in amplitude. F0 variability within an utterance was reduced under workload, suggesting that these utterances were produced with a flatter and perhaps less expressive F0 contour. Overall phrase durations and segmental durations were also reduced under workload, suggesting an overall increase in speaking rate as workload increased.

The patterns reported here are tendencies that emerged across a small number of subjects. Some differences were not always present for each subject. We believe that these patterns may be more consistent in an actual "high workload" environment than could be seen in this investigation in which performing poorly on the workload task had only minor consequences (compare, for example, Williams & Stevens, 1968, analysis of F0 characteristics in tape-recordings of actual conversations between pilots and flight controllers versus Hecker et al.'s (1969) analysis of F0 in a laboratory task designed to increase workload). Of the five speakers examined here, subject MC performed the JEX task at the highest level of difficulty and may have been the most highly motivated of the five subjects. It is interesting to note that, in general, MC showed the most consistent effects of workload on the acoustic-phonetic properties of speech.

The absence of any effect of workload on formant frequencies in combination with the other findings suggests that the main effect of workload occured at or below the level of the larynx. The changes in amplitude, F0 characteristics, and spectral tilt that we found in this study may be related to changes in the shape and variability of the glottal waveform (Hecker et al, 1968). We are currently analyzing our data further to determine the extent to which the various acoustic changes described above can be ascribed to this one source.

In summary, the results of this study demonstrate a number of reliable changes in the acoustic-phonetic properties of speech produced under increased cognitive workload. The findings add to a growing body of literature showing that talkers will consistently modify their speech in response to both physical and mental changes in their immediate environments. These results have important implications for the use of speech recognition devices in severe environments.

References

Brenner, M., Branscomb, H. H., and Schwarz G. E. (1979). Psychological stress evaluator-Two tests of a vocal measure. *Psychophysiology*, **16**, 351-357.

Goldman Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Hansen, J. H. L. (1988). Analysis and compensation of stressed and noisy speech with application to robust automatic recognition. Unpublished doctoral dissertatioin. Georgia Institute of Technology.

Hecker, M. H. L., Stevens, K. N., von Bismarck, G., and Williams, C. E. (1970). Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, **44**, 993-1001.

Kuroda, I., Fujiwara, O., Okamura, N., and Utsuki, N. (1976). Method for determining pilot stress through analysis of voice communication *Aviation, Space, and Environmental Medicine*, **47**, 528-533.

Jex, H. R., McDonnell, J. D., and Phatak, A. V. (1966). A 'critical' tracking task for manual control research. In Moray, N., editor, *Mental Workload*. New York: Plenum Press.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In Valdman, A., editor, *Emotions in personality and psychopathology*, pp. 493-529. New York: Academic Press.

Streeter, L. A., MacDonald, N. H., Apple, W., Krause, R. M., and Galotti, K. M. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, **73**, 1354-1360.

Tolkmitt, F. J., and Scherer, K. R. (1986) Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology*, **12**, 302-312.

Williams, C. E., and Stevens, K. N. (1969) On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, **40**, 1369-1372.

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 15 (1989)
Indiana University

Current Computer Facilities in the Speech Research Laboratory [1]

Robert H. Bernacki, Dennis M. Feaster,
Luis R. Hernandez, and Jerry C. Forshee

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] **Equipment and software development was supported, in part, by NIH Research Grant DC-00111-13, in part, by NSF Research Grant IRI-86-17847, and, in part, by USAF AAMRL Contract NO. AF-F-33615-86-C0549 to Indiana University in Bloomington, IN.**

Abstract

This report describes the expansion and development of the computer facilities and software in the Speech Research Laboratory for the period from 1985 through 1989.

Current Computer Facilities in the Speech Research Laboratory

Throughout the history of the Speech Research Laboratory at Indiana University, there has been a major commitment of resouces for the development of a facility capable of supporting virtually every aspect of speech research. In a continuing effort to provide researchers with the state-of-the-art acoustic-phonetic analysis and synthesis tools, we have expanded the computer facilities from the previous central computer and terminal environment to one using several networked graphics workstations. As described in our most recent instrumentation review (Forshee, 1984), the laboratory facilities consisted of a VAX 11/750 with a DSC 200 analog interface system for signal processing and general purpose computing, a Symbolics Lisp machine for symbolic computing, and three PDP 11/34 systems for online experimental control and data collection.

Three areas of expansion of computing resources and connectivity were completed this year. The first was the acquisition of several VAX workstations. Two desktop color VAXstation 3100s, and a VAXlab 3500 operating under VMS were installed to serve as dedicated speech processing and stimulus preparation workstations. The DSC 200 analog interface system was transferred from the VAX 11/750 to the VAXlab 3500 to serve the VAXstation 3100s.

The second area of expansion was to the laboratory networking system. This included an expansion of Ethernet to include the two VAXstations and the VAXlab, two IBM PC compatibles and two Apple MacIntosh IIs. Laboratory connectivity to the campus Ethernet has been made by the addition of two terminal servers and an Ethernet bridge. The campus system additionally provides a link to Internet and BITNET.

The third expansion was in the area of manuscript preparation. We acquired two MAC-IIs with Ethernet interfaces, an Apple Laserwriter, and upgraded two IBM clones to AT compatibility. We are also planning to connect the PDP 11/34 systems to the network in the near future. The VAX 11/750 now provides personal accounts for wordprocessing, EMAIL, and statistical analyses, as well as LEXICON accounts for computational analyses (Luce, 1986).

The following sections present brief details of these hardware systems and our software development efforts. Additional information on specific software packages may be available in the form of unpublished documentation.

The Hardware

VAX Hardware. The microVAX expansion centered around a VAXlab 3500 system. This system was configured with a 19" color monitor, 16 Mbyte of memory, an Ethernet port, a SCSI port, a mouse, a TK70 cartridge tape drive, and a CDC 9720 850Mbyte hard disk with a QD34 controller. This VAXlab serves as the hub of a VAX cluster including two VAXstation 3100 workstations. These computers were each configured with a 104Mbyte hard disk, a 19" color monitor, 16 Mbyte memory, an Ethernet port, a SCSI port, and a mouse. The VAXstations will be upgraded with additional 330 Mbyte hard disks during the first quarter of 1990. The three computers were configured in a DEC LAVC (Local Area VAX Cluster), with the VAXlab serving as the boot node, and the VAXstations as satellite nodes. In addition to the new VAX workstations, the VAX 11/750 was upgraded with a DEC TU81-plus High Density Magnetic tape drive (9 track, 6250 bpi). This unit provides an efficient mass-storage backup capability for all systems on the local DECnet.

DSC 200 Analog I/O System. The transfer of the DSC 200 system from the VAX 11/750 required the installation of a QBUS-to-Unibus converter on the VAXlab 3500. The DSC 200 analog interface system was located with the VAXlab adjacent to an IAC booth for recording and execution of single-subject experiments. In addition to the transfer to the VAXlab, the DSC 200 system was upgraded with filters for 16 kHz sampling rates to provide compatibility with other laboratories and public speech databases using this format. The subject interface in the IAC booth consists of a terminal and headphones with attached microphone. The VAX 3100 workstations located in other rooms utilize the DSC 200 device on the VAX 3500 via remote DSC 240 analog interface units. In addition to the computer upgrades, new analog audio acquisitions include a NAD stereo cassette deck model 6155 and a Nakamichi DMP-100 Digital Mastering Processor with a Sony SLD-420 BETA recorder. These units are interfaced with the DSC 200 system and provide media compatibility as well as high quality audio archiving. For the measurement of the glottal function, a special microphone apparatus (Sondhi, 1975) was constructed and installed for use with the DSC 200 in the VAXlab IAC booth.

DECnet Expansion. The availability of connectivity is central to the implementation of a multiple workstation environment. In addition to local networking, we made provisions for connections to the campus network and via that system of connections to the national INTERNET. Our local laboratory network connects to the Psychology building baseband Ethernet via a DEC LAN Bridge 100. From there, a DEC DELNI interconnect distributes the thickwire Ethernet to the VAX 11/750, the Symbolics, the two DEC terminal servers, and to a DEC DEMPR thinwire Ethernet repeater. The thinwire Ethernet serves the VAXlab, the two VAXstations, the PDP 11/34s, the IBM PC/AT compatibles, and the MAC IIs. Ethernet interface cards were installed on the IBM PC/AT compatibles and the the MAC IIs to provide thinwire connectivity.

In order to provide greater flexibility for terminals on the network, we installed two DECserver terminal servers, providing a total of sixteen terminal lines. The DECservers allow users to log simultaneously into multiple computers on the local or campus networks. Benefits of the terminal servers include reduced overhead on the VAX 11/750 and higher speed throughput to terminals. For offnet communications, two 2400 baud modems were installed on a DECserver.

Mini/Micro Computers. For manuscript preparation, two Apple MAC-II personal computers, in conjunction with an Apple Laserwriter and Laserwriter Nt (Postscript) board were installed for word-processing and graphics. As mentioned above, these computers are connected to the laboratory Ethernet and are supported by both DECnet and TCP/IP software.

To provide file transfer compatibility, all three PDP-11/34 minicomputers and the VAX 11/780 now have Kennedy Inc. magnetic tape drives. We have 9300 series (9 track, 125ips) and 9100 series (9track, 75ips) drives. Other upgrades to these mincomputers consisted of replacing the previous obsolete hard drives with CDC 9720 368Mbyte hard disks and Emulex UD33 disk controllers. The additional mass storage capability greatly facilitates scheduling of behavioral experiments and setup efficiency by permitting several digital stimulus databases to reside online simultaneously.

The Software

VAXstation Software (External). To provide a broad spectrum of connectivity, TCP/IP was installed on the VAX computers. In addition to local file transfers, TCP/IP provides access to computing resources at remote laboratories via TELNET on INTERNET. The VAX workstation complement of system software includes VWS, DECWindows, FORTRAN compiler, C compiler, and LAVC software under VMS.

Extending our utilization of Signal Technology's ILS signal processing environment, the VAX 11/750 license was transferred to the VAXlab and the package upgraded to ILS version 6.1. This version offers an improved spectrographic display and a more intuitive windowed environment. For the creation of experimental stimuli, a recent version of the Klatt software synthesizer was installed on the VAXlab. Following our philosophy of maintaining compatibility with systems in various other speech laboratories, we also installed the SPEED software designed by Phil Rubin of Haskins Laboratories. This program provides waveform editing and speech analysis in a flexible windowed environment on the DEC VAXlabs and is now used by several speechlaboratories.

VAXstation Software (In-house). In order to take full advantage of the new hardware capabilities, we began the process of updating and porting our various software packages from the previous graphics terminal and minicomputer environment to the VAXcluster workstation environment. Among the programs ported from the PDP 11/34 RT-11 environment were WAVMOD (Bernacki, 1981) and a new version of WAVES (Luce and Carrell, 1981) now renamed KiWE (Keypad Waveform Editor).

While ILS provides a broad range of analysis capabilities, the early versions were rather awkward for large scale acoustic-phonetic analysis. To provide a more user-friendly and custom-tailored environment, we created a program called SRD (Speech Read). This ILS compatible program provides an integrated display, query, and measurement environment for time and spectral domain data as well as phonetic labeling. Within the program, users can mark acoustic-phonetic events of interest in the signal. Marked segments or single frame data can then be labeled with any of a variety of phonetic systems. Optionally, the labeled acoustic data can be recorded to a parameter file. A non-interactive version of SRD, SPP (Speech Post-Processor), utilizes previously segmented SRD files to apply new analysis measures post-hoc to the speech waveforms. A companion program, TSA (Talker Specific Analysis), permits convenient statistics and histograms for the "fine-tuning" of the analysis programs. The application of TSA ensures optimal speech analysis information for use in SRD. Another companion program, SDP (Speech Data Processor), combines parameter files for all tokens across conditions and words and formats the data for statistical analyses using BMDP. The SRD program has become a mainstay of acoustic-phonetic analysis in the laboratory.

The first version of SRD was implemented on the VAX 11/750 and a Tektronix 4027 compatible graphics terminal. The displays were fixed on the screen and limited to waveform, enhanced pseudo-spectrogram, fundamental frequency, energy, cross-section spectrum, and a vowel-space plot. SRD was transferred to the VAXlab and expanded to take advantage of the DEC VWS operating system. This new version, WSRD, permits the segmenting task to be performed at a quicker pace, due to increased computation speed and improved user interface. WSRD displays are presented in windows which may be re-sized, positioned and shaded according to user preferences. The entire layout of the display can be custom tailored. With the availability of greater monitor resolution, more displays are available at one time, and the editing process becomes much more precise. The program provides mouse control and menu driven commands.

The SAP (Speech Acquisition Program) developed by Dedina (1987) was transferred to the VAXlab for the online sampling of speech. Digitizing may be performed under benign conditions or during experimental manipulations of the speaking environment. SAP provides visual cues on a terminal while digitizing an utterance into a file. This process is carried out via a control/configuration file. Hundreds of utterances may be collected automatically during one run of the program. Ongoing software development of SAP provides improved stimulus presentation and user interface.

An expanded version of the SAP software also has been developed. This program, called IOSAP, provides for auditory stimulus presentation from digitized files while simultaneously digitizing utterances of a talker in an on-line experiment. This will enable us to implement an on-line NAMING paradigm, in which a subject responds vocally to an auditory signal. The start of the digitized recording is time-aligned to the digitized audio output. This permits reaction-time measures to be made from the utterance to within one millisecond accuracy.

A special utility program, RTSTM, for transferring files between RT-11 and VMS via magnetic tape was written in response to the difficulties in applying DEC utility software to this task. RTSTM provides for wildcard file specification and file count for transfer control. In addition, the software preserves RT-11 file names during transfer to VMS.

Due to the complexity and variety of systems and software tools available to researchers in our laboratory, we developed a help utility called SRLHELP. This utility provides convenient access to information on all Speech Research Laboratory resources. It covers basic procedures such as backups, file transfering, and use of software packages, as well as more advanced procedures. It is implemented as a menu driven package with a versatile text editor style of search and browse. The utility is implemented as a program operating on various text databases, thus providing for rapid updates. The system has been very successful in providing a wide range of information while eliminating hardcopy documentation.

Manuscript Preparation. The Apple MAC-II personal computers, in conjunction with a Laserwriter, are now utilized for word-processing in Postscript. Data may be transferred from the VAX computers to the MAC-IIs via DECnet and Alisa System's TSSnet software. Data may then be formatted and plotted with packages such as Excel and Cricket Graph. Hardcopy of plots may then be obtained from the Laserwriter on either paper or overhead transparencies. The PC/AT workstations provide media transfer compatibility as well as supporting word processing, the Microsoft CHART plotting package, and the Lotus 1-2-3 package for spreadsheet analysis.

Symbolics Software. Software upgrades for the SYMBOLICS Lisp machine have been installed for versions of the MIT programs ALEXIS, SPIRE, and SEARCH, providing an integrated environment for lexical searching, speech segmentation, and acoustic analysis. To support ALEXIS, we converted our SCRL dictionary from a VMS format to one suitable for the SYMBOLICS (Luce, 1986). For waveform transfer from the VMS environment, we created a file conversion program to provide SPIRE with access to ILS speech files residing on the VAX computers. The file transfer takes place over the network via DECnet DNA 6.0 software.

Summary

In summary, as the research demands have changed over the years, we have had to develop new and more sophisticated software and hardware systems to meet the needs of the researchers in our laboratory. Changes in the computer resources have been continuous in an environment like ours at Indiana where new ideas and techniques are constantly being generated to solve a variety of problems in speech perception, analysis, synthesis, and spoken word recognition.

References

Bernacki, B. (1981). WAVMOD: A program to modify digital waveforms. *Research on speech perception progress report no. 13.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Dedina, M. (1987). SAP: A speech acquisition program for the SRL-VAX. *Research on speech perception progress report no. 13.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Forshee, J. C. and Nusbaum, H. C. (1984). An update on computer facilities in the speech research laboratory. *Research on speech perception progress report no. 10.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P. A. and Carrell, T. D. (1981). Creating and editing waveforms using WAVES. *Research on speech perception progress report no. 7.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P. A. (1986). Neighborhood of words in the mental lexicon. *Research on speech perception technical report no. 6.* Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Sondhi, M. M. (1975). Measurement of the glottal waveform. *Journal of the Acoustical Society of America*, **57**, 228-232.

IV. PUBLICATIONS

Papers Published

Charles-Luce, J. & Luce, P.A. (1989). Some structural characteristics of young children's lexicons. *Journal of Child Language*, **17**, 205-215.

Davis, S. (1989). Cross-vowel phonotactic constraints. *Computational Linguistics*, **15**, 109-110.

Davis, S. (1989). On a non-argument for the rhyme. *Journal of Linguistics*, **25**, 211-217.

Davis, S. (1989). The location of [continuant] in feature geometry. *Lingua*, **78**, 1-22.

Davis, S. & Tsujimura, N. (1989). The morphophonemics of Japanese verbal conjugation: An autosegmental account. *Proceedings of the Fifth Eastern States Conference on Linguistics* (pp. 488-499).

Davis, S. & Summers, W. V. (1989). Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics*, **17**, 339-353.

Gierut, J.A. (1989). Maximal opposition approach to phonological treatment. *Journal of Speech and Hearing Disorders*, **54**, 9-19.

Gierut, J.A. (1989). Developing descriptions of phonological systems: A surrebuttal. *Applied Psycholinguistics*, **10**, 469-473.

Gierut, J.A. & Pisoni, D.B. (1989). Speech perception. In J. Northern (Ed.), *Study guide for handbook of speech-language pathology and audiology* (pp. 60-68). Philadelphia: B.C. Decker.

Gierut, J.A., Elbert, M. & Dinnsen, D.A. (1989). Issues of linguistic analysis and experimental design: Reply to Diedrich. *Journal of Speech and Hearing Research*, **32**, 219-222.

Goldinger, S.D., Luce, P.A. & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Johnson, K. (1989). Higher formant normalization results from auditory integration of F2 and F3. *Perception & Psychophysics*, **46**, 174-180.

Logan, J.S., Greene, B.G. & Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566-581.

Martin, C.S., Mullennix, J.W., Pisoni, D.B. & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 676-684.

Mullennix, J.W. & Pisoni, D.B. (1989). Speech perception: Analysis of biologically significant signals. In R.J. Dooling and S.H. Hulse (Eds.) *The comparative psychology of audition: Perceiving complex sounds* (pp. 97-128). Hillsdale, NJ: Erlbaum.

Mullennix, J.W., Pisoni, D.B. & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Ozawa, K. & Logan, J. S. (1989). Perceptual evaluation of two speech coding methods by native and non-native speakers of English. *Computer Speech and Language*, **3**, 53-59.

Pisoni, D.B. & Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, **13**, 577-587.

Stemberger, J.P. (1989). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In B.Baars (Ed.), *The psychology of errors: A window on the mind?* New York, NY: Plenum.

Summers, W.V., Johnson, K., Pisoni, D. & Bernacki, R. (1989) An addendum to "Effects of noise on speech production: Acoustic and perceptual analyses" [J.Acoust.Soc.Am. 84, 917-928 (1988)], *Journal of the Acoustical Society of America*, **86**, 1717-1721.

Manuscripts Accepted for Publication (In Press):

Charles-Luce, J. (in press). The effects of semantic context on voicing neutralization. *Phonetica*

Charles-Luce, J., Luce, P.A., & Cluff, M. (in press). Retroactive influence of syllable neighborhoods. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Connine, C.M. & Mullennix, J.W. (in press). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Davis, S. (in press). An argument for the underspecification of coronal in English. *Linguistic Inquiry* **21**.

Davis, S. (in press). Coronals and the phonotactics of nonadjacent consonants in English. In C. Paradis & J.-F. Prunet (Eds.), *The special status of coronals.*

Dedina, M.J. & Nusbaum, H.C. (in press). PRONOUNCE: A program for pronunciation of new words by analogy. *Computer Speech and Language.*

Johnson, K. (in press). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America.*

Johnson, K. (in press). Contrast and normalization in vowel perception. *Journal of Phonetics.*

Luce, P.A., Pisoni, D.B. & Goldinger, S.D. (in press). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Pisoni, D.B. (in press). Modes of processing speech and nonspeech signals. In I.G. Mattingly (Ed.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum.

Pisoni, D.B., Logan, J.S. & Lively, S.E. (in press). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In H. Nusbaum and J. Goodman (Eds.), *Development of speech perception: The transition from recognizing speech sounds to spoken words.* Cambridge: MIT Press.

Ralston, J.V., Pisoni, D.B. & Mullennix, J.W. (in press). Comprehension of synthetic speech produced by rule. In R. Bennet, A. Syrdal and S. Greenspan (Eds.), *Behavioral aspects of speech technology: Theory and applications.* New York: Elsevier.

V. Speech Research Laboratory Staff, Faculty, and Technical Personnel
(1/1/89 - 12/31/89)

Research Personnel:

David B. Pisoni, Ph.D.	Professor of Psychology and Director

Theodore S. Bell, Ph.D.	Visiting Asst. Professor of Psychology[1]
Beth G. Greene, Ph.D.	Research Scientist and Assoc. Director[2]
John W. Mullennix, Ph.D.	Research Associate[3]
James V. Ralston, Ph.D.	Visiting Asst. Professor of Psychology
W. Van Summers, Ph.D.	Research Associate[4]

Dawn M. Behne, Ph.D.	NIH Post-doctoral Trainee
Stuart A. Davis, Ph.D.	NIH Post-doctoral Trainee[5]
Keith Johnson, Ph.D.	NIH Post-doctoral Trainee

Michael S. Cluff, B.S.	NIH Pre-doctoral Trainee

Stephen D. Goldinger, B.A.	Graduate Research Assistant
Mary Jo Lewellen, M.A.	Graduate Research Assistant
Nancy L. Lightfoot, B.A.	Graduate Research Assistant
Scott E. Lively, B.S.	Graduate Research Assistant
John S. Logan, B.S.	Graduate Research Assistant
Heng Jie Ma, B.S.E.E.	Graduate Research Assistant
Joanne K. Marcario, B.S.	Graduate Research Assistant
Christopher S. Martin, B.A.	Graduate Research Assistant
Brigette R. Oliver, B.A.	Graduate Research Assistant
Michael A. Stokes, B.A.	Graduate Research Assistant

---

[1] On leave from the UCLA School of Medicine, Department of Surgery, Division of Head and Neck, Los Angeles, CA 90024.

[2] Also, Center for Reading and Language Studies, School of Education, Indiana University, Bloomington, IN 47405.

[3] Now at Department of Psychology, Wayne State University, Detroit, MI 48202

[4] Now at Army Audiology and Speech Center, Walter Reed Army Center, Washington, DC 20307.

[5] Now at Linguistics Department, Indiana University, Bloomington, IN 47405

Technical Support Personnel:

Robert H. Bernacki, B.A.	Research Analyst
Cheryl L. Blackerby	Administrative Secretary
Dennis M. Feaster, B.S.	Software Development Specialist
Jerry C. Forshee, M.A.	Computer Systems Analyst
Luis R. Hernandez, B.A.	Software Development Specialist
David A. Link	Electronics Engineer
Gary Link	Technical Assistant

Denise R. Beike	Undergraduate Research Assistant
Amy B. Lawlor	Undergraduate Research Assistant